

AD-A063 181

POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 9/3  
PROGRESS REPORT NUMBER 43 TO THE JOINT SERVICES TECHNICAL ADVIS--ETC(U)  
NOV 78 A A OLINER F44620-78-C-0074  
POLY-MRI-452.43-78

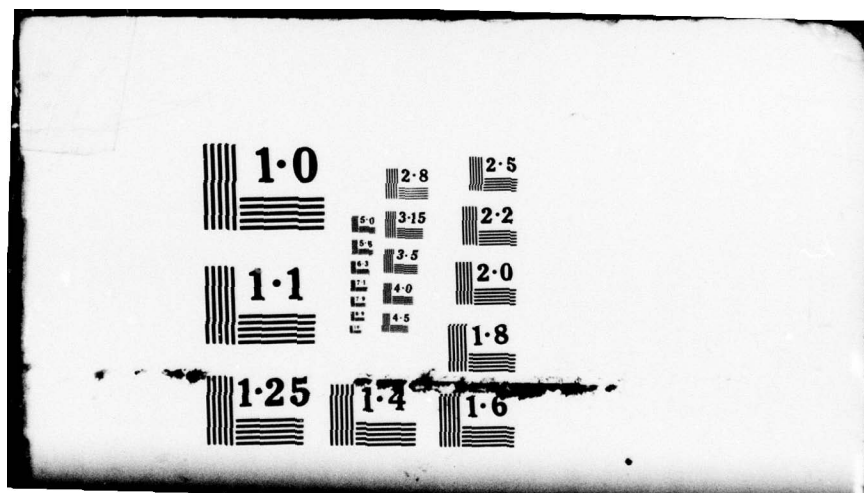
UNCLASSIFIED

NL

1 OF 6  
ADA  
063181

12/1/78





AD A0 63 181

**LEVEL**

A050265

#42

**Progress Report No. 43**

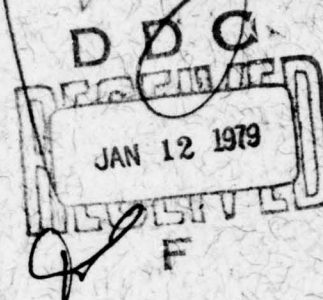
15 September 1977 through 14 September 1978

**The Joint Services  
Technical Advisory Committee**

**Representing:** The Air Force Office of Scientific Research  
The U.S. Army Research Office  
The Office of Naval Research

**Submitted by:** Arthur A. Oliner, Director  
Microwave Research Institute

**Coordinated by:** Jean Beirne Maher



**DDC FILE COPY**

**Report**  
R-452.43-78

**Submitted**  
**November**  
**1978**

for  
Contract F44620-78-C-0074  
Project No. 4751

This document has been  
approved for public release and  
sale; its distribution is unlimited.

**Polytechnic**

Polytechnic Institute of New York  
333 Jay Street, Brooklyn, New York 11201

Qualified requestors may obtain additional copies from the Defense Documentation Center; all others should apply to the Clearinghouse for Federal Scientific and Technical Information

79 01 12 013

## **Progress Report No. 43**

15 September 1977 through 14 September 1978

### **The Joint Services Technical Advisory Committee**

**Representing:** The Air Force Office of Scientific Research  
The U.S. Army Research Office  
The Office of Naval Research

**Submitted by:** Arthur A. Oliner, Director  
Microwave Research Institute

**Coordinated by:** Jean Beirne Maher

**Report  
R-452.43-78**

**Submitted  
November  
1978**

**for  
Contract F44620-78-C-0074  
Project No. 4751**

**This document has been  
approved for public release and  
sale; its distribution is unlimited.**

# **Polytechnic**

**Polytechnic Institute of New York  
333 Jay Street, Brooklyn, New York 11201**

Qualified requestors may obtain additional copies from the Defense Documentation Center; all others should apply to the Clearinghouse for Federal Scientific and Technical Information

## PREFACE

This report to the Joint Services Technical Advisory Committee presents a summary of the research programs in the broad field of electronics conducted during the past year at the Polytechnic Institute of New York. These programs are pursued within the framework of the Microwave Research Institute, and they involve the academic research activities of faculty in the departments of Electrical Engineering, Physics and Chemistry. The research projects cover a broad spectrum ranging from basic theoretical investigations in physics, applied mathematics, and engineering to experimental efforts involving basic measurements and the development of devices and materials.

The format of this annual report permits a coherent presentation of the various phases of the Joint Services Electronics Program (JSEP) at the Polytechnic and their relation to ongoing research in electronics sponsored by other agencies. This presentation is intended for the information of the Air Force Office of Scientific Research, the Army Research Office and the Office of Naval Research and, in addition, the other sponsors who are individually acknowledged throughout the report. The principal aims of the JSEP are to initiate deserving lines of research in a timely fashion and to develop investigations to a stature sufficient to attract individual support on their own merits.

In the early days of the Microwave Research Institute, the research program consisted primarily of projects involving electromagnetics and microwave components. Although the name of the Institute has remained the same, the nature of the research programs has broadened substantially, and the programs now encompass a wide range of topics within the field of electronics. The current programs are organized into twelve areas: electromagnetics; acoustics; optics; quantum electronics; solid state and materials; wave-matter interactions; electric power engineering; communications; computers and computer-communication networks; safety, reliability and software engineering; systems, control and networks; and image processing.

Arthur A. Oliner  
Director

## TABLE OF CONTENTS

Preface .....	iii
Participating Faculty and Research Staff.....	x

### I. ELECTROPHYSICS

#### A. ELECTROMAGNETICS

1. High Frequency Fields Excited by a Line Source Located on a Concave Cylindrical Impedance Surface T. Ishihara and L. B. Felsen.....	2
2. High Frequency Surface Fields Excited by a Point Source on a Concave Perfectly Conducting Cylindrical Boundary L. B. Felsen and T. Ishihara.....	17
3. Periodic Structure GTD for Analysis of Mutual Coupling in Arrays on Concave Surfaces H. Ahn and A. Hessel.....	32
4. Limitations on Gain-Versus-Scan for a Dome Antenna H. Steyskal, A. Hessel and J. Shmoys.....	43
5. Blazed Diffraction Gratings for Frequency Scanned Antennas A. Hessel, J. Shmoys and S. T. Peng.....	47
6. Properties of the Shadow Cast by a Half-Screen When Illuminated by a Gaussian Beam A. C. Green, H. L. Bertoni and L. B. Felsen.....	54
7. Scattering of Surface Waves by a Dielectric Step Discontinuity: Oblique Incidence Case J. P. Hsu, S. T. Peng and A. A. Oliner.....	68
8. New Propagation Effects for the Inverted Strip Dielectric Waveguide for Millimeter Waves A. A. Oliner, S. T. Peng and J. P. Hsu.....	74
9. Interactive Mechanisms and Effects of Low-Level Millimeter Waves on Living Systems S. Motzkin, S. W. Rosenthal, L. Birenbaum, R. Melnick, C. Rubenstein, R. Remilly and S. Davidow.....	79

#### B. ACOUSTICS

1. Acoustoelectric Real Time Correlator L. Rosenheck, H. Schachter and W-C. Wang.....	89
2. An Acoustoelectric FM Demodulator H. Schachter, W-C. Wang, F. A. Cassara and L. Rosenheck.....	95

#### C. OPTICS

1. A Simple Criterion for Predicting Leaky Waves on Rib Waveguides for Integrated Optics S. T. Peng and A. A. Oliner.....	100
--	-----

## CONTENTS

2.	Modulation Sensitivity of Electro-Optic Slab Waveguides S. T. Peng .....	113
3.	Ray Analysis of Unstable Resonators S.H. Cho and L.B. Felsen.....	116
4.	Small Misalignment Effects in Unstable Resonators C. Santana and L.B. Felsen .....	121
5.	Propagation in Inhomogeneous Slab Waveguides -- Exact Solutions E. Navon and L.B. Felsen.....	123
6.	Guided Modes in a Graded Index Optical Fiber Surrounded by a Homogeneous Cladding E. Navon and L.B. Felsen.....	136
7.	The Diffraction of Gaussian Beams by Periodic Layers R.S. Chu, J.A. Kong and T. Tamir.....	143
8.	Bragg-Reflection Approach for Blazed Dielectric Gratings K.C. Chang and T. Tamir .....	150
9.	Effects of Phase Variations on Optical Parametric Mode- Coupling Dynamics E.S. Cassedy and M. Jain .....	156

## D. QUANTUM ELECTRONICS

1.	Theory of the Integrating Sphere for Pulsed Light Sources K. Park and W. T. Walter .....	164
2.	Two-photon Correlations in Amplified Laser Light G.D. Blake and D.B. Scarl .....	174

## E. SOLID STATE AND MATERIALS

1.	Metallic Field Effect at Free Metal Surfaces H.J. Juretschke, E. Segredo and M. Eschwei .....	182
2.	Charge-Induced Surface Stresses at Metal-Insulator Interfaces H.J. Juretschke, R. Boucarut and E. Segredo .....	185
3.	Elastoresistance and Electron Tunneling in Oxidized Iron T. Pignataro, H.J. Juretschke and M. Eschwei.....	188
4.	Solid State and Materials E. Banks, S. Nakajima, R. Sacks and M. Shone.....	191
5.	Ultrasonic and Microwave Dielectric Relaxation of Liquid Dialkyl Carbonates D. Saar, J. Brauner, H. Farber and S. Petrucci.....	195
6.	Dielectric Relaxation of Some 1:1 Electrolytes in Tetrahydrofuran and Diethylcarbonate D. Saar, J. Brauner, H. Farber and S. Petrucci.....	212

## CONTENTS

### F. WAVE-MATTER INTERACTIONS

1.	Classical Wave-Particle Duality N. Marcuvitz.....	225
2.	Wavepackets as Quasiparticle Systems N. Marcuvitz.....	234
3.	Renormalization of Maxwell's Equations for Turbulent Plasmas S. Barone and N. Marcuvitz.....	241
4.	A New Approach to Some Nonlinear Ionospheric Plasma Turbulence Problems S. Barone.....	246
5.	Preliminary Report of Numerical Simulation of Type II Irregularities in the Equatorial Electrojet S. Barone, N. Marcuvitz, R. Pascone and N. Solimene.....	249
6.	Nonlinear Theory of Type II Irregularities in the Equatorial Electrojet S. Barone.....	266
7.	High Power Microwave Propagation Through the Atmosphere N. Marcuvitz and N. Solimene.....	272
8.	Kinetic Theory Treatment of Metal Evaporation Front M. Newstein, N. Solimene and J. Hammer.....	285
9.	Functional Equation Approach to the Three-Wave Nonlinear Interaction S. T. Peng and E. S. Cassedy.....	291
10.	Analysis of Strong Electromagnetically Induced Spherically Imploding Shocks Y. Fujimoto and E. A. Mishkin.....	298

### G. ELECTRIC POWER ENGINEERING

1.	Development of the Design for Iron-Cored Synchronously Operating Linear Motors E. Levi.....	305
2.	Computer-Aided Conformal Mapping of Magnetic Fluxes in Saturated Inductor Motors E. Levi, J. P. Lee, F. Lalezari and M. Gemelos.....	311
3.	Modal Representation of E-M Fields in Laminated Moving Conductive Ferromagnetic Media B. R. Cheo, E. Levi and K. C. Chang.....	315

ACCESSION for	
NTIS	W. E. Section <input checked="" type="checkbox"/>
DDC	E. E. Section <input type="checkbox"/>
UNANNOUNCED	
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY NOTES	
Dist.	GIAL
<div style="font-size: 2em; font-weight: bold; position: absolute; left: 10px; bottom: 10px;">A</div>	

## CONTENTS

### II. SYSTEMS

#### A. COMMUNICATIONS

1.	Comparison of Sequential Partition Detectors (SPD) with Other Nonparametric Sequential Detectors R. F. Dwyer and L. Kurz .....	324
2.	A Comparison of Parametric and Partition Sequential Detectors in Radar and Sonar Problems R. F. Dwyer and L. Kurz .....	330
3.	Recursive Equalization for Timing Jitter Suppression in Partial Response Systems M. Kavehrad and L. Kurz .....	335
4.	Robust Score Estimation for Simple Linear Rank Detectors I. M. Habib and L. Kurz .....	339
5.	Extrapolation of Band-limited Signals Using a Stochastic Approximation Algorithm I. Kadar and L. Kurz .....	344
6.	Influence of Nonlinearities on the Performance of Partial Response System M. Kavehrad and L. Kurz .....	348
7.	Robustized Sequential Partition Detectors H. S. Ashtiani and L. Kurz .....	362
8.	A Representation Theory and Its Applications R. Chassaing and L. Kurz .....	372
9.	A Uniform Power Spectral Density Jamming Signal F. A. Cassara .....	382
10.	Transient Acquisition Behavior of the Cross-Coupled Phase-Locked Loop FM Demodulator F. A. Cassara, H. Schachter and G. Simowitz .....	387

#### B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

1.	Second Order Greedy Algorithms for Centralized Teleprocessing Network Design A. Kershenbaum and R. Boorstyn .....	391
2.	Adaptive Routing in Networks R. Boorstyn and A. Livne .....	402
3.	Modeling of Digital Systems Using Microprocessors D. R. Kaufman and E. J. Smith .....	405
4.	Adaptive Channel Capacity Controllers for Communication Networks L. Shaw and K. Sohraby .....	416

## CONTENTS

### C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

1.	Software Modeling Studies	
	M. L. Shooman and H. Ruston .....	423
	Application of Models to Software Engineering .....	423
	Software Error, Reliability, and Availability Models.....	425
	Test Models and Techniques.....	428
	Complexity Models.....	432
	Other Modeling Research in Progress.....	435

### D. SYSTEMS, CONTROL AND NETWORKS

1.	Notes on n-dimensional System Theory	
	D. C. Youla and G. Gnani .....	449
2.	The Identification of Linear Dynamical Systems from Time-Domain Measurements: A Critical Study of Certain Aspects of Prony's Method	
	D. C. Youla .....	466
3.	On Order Determination of Linear AR Models	
	F. Nakajima and F. Kozin .....	488
4.	Characterization of Consistent Estimates	
	F. Nakajima and F. Kozin .....	492
5.	A Nonlinear Servo for Linear Systems	
	L. Shaw and H. Gambe .....	496
6.	Efficient Discrete Fourier Transformation of Real Vectors	
	T. W. Parsons.....	504
7.	Some Comments on Improving the Efficiency of Least Squares Estimation	
	I. Kadar and L. Kurz .....	508

### E. IMAGE PROCESSING

1.	Adaptive Frequency Domain Estimators	
	A. Papoulis .....	510
2.	Bound on the Probability of Misclassification of Two-Dimensional Images	
	L. Kurz and P. Legakis .....	527
3.	Quadratic Tests in M-Gray Level Detection	
	L. Kurz and P. Legakis .....	535
4.	Robustized Vector Form of Gladyshev's Theorem with Application to Estimation of Parameters in a Linear Model	
	I. Kadar and L. Kurz .....	541
	Publications and Reports.....	547

POLYTECHNIC INSTITUTE OF NEW YORK

G. Bugliarello, President  
A. Gold, Provost  
R.J. Cresci, Associate Provost for Research

B.J. Bulkin, Dean of Arts and Sciences

A.B. Giordano, Dean of  
Graduate Studies and Engineering

MICROWAVE RESEARCH INSTITUTE DIRECTORATE

A.A. Oliner, Director

S.W. Rosenthal, Assistant Director

J.B. Maher, Executive Assistant

FACULTY AND RESEARCH STAFF PARTICIPATING IN  
MICROWAVE RESEARCH INSTITUTE PROGRAMS

DEPARTMENT OF CHEMISTRY

Faculty

E.M. Pearce, Department Head

E. Banks

P. Hoggard

N.C. Peterson

S. Petrucci

Postdoctoral Fellows

R. Carvalho

C. Kosky

S. Nakajima

Fellows, Trainees and Research Associates

M. Shone

R. Sachs

DEPARTMENT OF ELECTRICAL ENGINEERING

Faculty

T. Tamir, Department Head

L.B. Bergstein

A.B. Giordano

N. Marcuvitz

H. Schachter

H.L. Bertoni

S.H. Gross

J. Mirza

B. Senitzky

J.J. Bongiorno, Jr.

A. Hessel

E. Mishkin

L. Shaw

R.R. Boorstyn

A. Kershenbaum

M.C. Newstein

J. Shmoys

F.A. Cassara, Jr.

A. Klappholz

A.A. Oliner

M.L. Shooman

E.S. Cassedy

F. Kozin

I. Palocz

L.M. Silber

B.R. Cheo

L. Kurz

A. Papoulis

E.J. Smith

D. Davids

A.E. Laemmel

S.W. Rosenthal

W.-C. Wang

R. Drenick

J.T. LaTourrette

N. Rubin

D.C. Youla

H. Farber

E. Levi

H. Ruston

Z. Zabar

L.B. Felsen

Research Staff

K.C. Chang

B. Rudner

N. Solimene

W.T. Walter

Research Faculty

L. Birenbaum

S.H. Cho

S.P. Kuo

T. Parsons

S.T. Peng

## DEPARTMENT OF ELECTRICAL ENGINEERING

### Postdoctoral Fellow

H. J. Eun

### Fellows, Graduate Assistants and Graduate Students

H. Ahn	J. Hammer	R. Mihailovic	V. Shah
K. Aupperle	C. Hechtman	S.K. Park	M.J. Shiau
H. Beca	T.I. Hsu	R. Pascone	C. Tsai
G. Borriello	H.M. Huang	B. Poole	B. Tomasic
J. Chan	K.Y. Huang	M. Post	M. Ulema
C.C. Chen	A. Kamel	B. Rabinowitz	E. Voudouri
W. Chuang	T.H. Kim	L. Rosenheck	C. C-H. Wang
J. Durnin	R. Kinasewitz	R. Roy	D.M.H. Wu
P. Einziger	K.K. Ko	C. Rubenstein	C. Yellin
R. Faaland	G. Konesky	A. Ruzycki	Y-M. Yu
H. Gambe	J. Lipka	V. Sahin	
J. Goldfinger	S. Mahfooz	A. Sanchez Corpas	

## DEPARTMENT OF PHYSICS

### Faculty

T. Kjeldaas, Department Head

H.J. Juretschke  
W. Kiszenick

D.C. Mattis  
B. Post

D.B. Scarl  
H. Schleuning

### Research Scientists

S.R. Barone

J. Gullardo

R. Pena

### Research Associate

M. Eschwei

### Fellows, Graduate Assistants and Graduate Students

H. Al-Salamah	J. Camacho	S. Lambrakos	E. Montvidas
L. Amani	P. Gong	J. Marincic	R. Pimpinella
Y. Amani	J. Hammer	K. Masand	F. Robbins
G. Blake	T. Horn	G. McCreary	P. Wang

## MICROWAVE RESEARCH INSTITUTE

### MRI Technical Reports

Editor:

J.B. Maher

Graphic Arts:

R. Gnaffo-Vastano

V. Goellner

Typing:

A. Drury

J.B. Maher

Composition:

T. Rowan

P. Testagrossa

## WORD PROCESSING CENTER

M. Fischetti, Supervisor

L. Babikian  
E. Cummings

C. Devlin  
R. Drucker

F. Kay-Kamara  
E. MacDonald

## I. ELECTROPHYSICS

- A. ELECTROMAGNETICS
- B. ACOUSTICS
- C. OPTICS
- D. QUANTUM ELECTRONICS
- E. SOLID STATE AND MATERIALS
- F. WAVE-MATTER INTERACTIONS
- G. ELECTRIC POWER ENGINEERING

# HIGH FREQUENCY FIELDS EXCITED BY A LINE SOURCE LOCATED ON A CONCAVE CYLINDRICAL IMPEDANCE SURFACE

T. Ishihara and L. B. Felsen

## A. Introduction

When high-frequency fields impinge on a concave surface with large radius of curvature, the induced fields thereon differ markedly from those for the well explored convex case. A recent study<sup>1</sup> has dealt in detail with the two-dimensional problem posed by line source excitation on the interior boundary of a perfectly conducting circular cylinder. Alternative field representations, accounting only for propagation phenomena on the surface segment between source and observation points, were formulated and evaluated asymptotically. The field expressions so obtained were interpretable physically in terms of constituents involving whispering gallery modes, ray-optical fields, near field effects, continuous spectrum and canonical integral contributions. Extensive numerical comparisons showed the accuracy and range of applicability of each. This investigation has provided fundamental physical and quantitative insight into the field behavior on the perfectly conducting boundary.

Here the analysis is extended to accommodate surfaces with non-vanishing surface impedance  $Z_s$ . Because the alternative field representations (Section B) for  $Z_s \neq 0$  are derived by the same techniques as those for  $Z_s = 0$ , the presentation is kept concise, with frequent reference made to the earlier study<sup>1</sup> for some of the details. Representative numerical calculations in Section C show how increasing surface impedance affects the surface field. When losses are appreciable, the closely bound whispering gallery modes are rapidly attenuated, leaving a properly formulated ray-optical field as the dominant and adequate contributor.

## B. Alternative Representations

### 1. Basic Green's Function

The relevant magnetic line source Green's function for the interior of a circular cylinder with radius  $a$  and constant surface impedance  $Z_s$  is formulated in an infinite angular space wherein the azimuthal coordinate  $\phi$  ranges from  $-\infty$  to  $+\infty$ .<sup>2</sup> With the source point located at  $(\rho', \phi')$  in a cylindrical  $(\rho, \phi)$  coordinate system, the Green's function satisfies the inhomogeneous wave equation

$$\left( \frac{1}{\rho} \frac{\partial}{\partial \rho} \rho \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \phi^2} + k^2 \right) G(\rho, \phi; \rho', \phi') = -\frac{1}{\rho} \delta(\phi - \phi') \delta(\rho - \rho') \quad (1)$$

with an "angular radiation condition" in the  $\phi$  direction, and the radial boundary condition

$$\frac{\partial}{\partial \rho} G = ikZ'G \text{ at } \rho = a ; \quad G \text{ finite at } \rho = 0 \quad (1a)$$

Here  $Z' = Z_s/Z_0$  is a normalized surface impedance, and  $Z_0$  is the impedance of free space. A time factor  $\exp(-i\omega t)$  is suppressed.

Expressing the Green's function  $G$  as a contour integral and removing diffraction effects at the origin arising from the "angularly matched" condition,<sup>2</sup> one obtains a modified form that contains only the essential propagation phenomena from the source point to the observation point in the presence of a cylindrical boundary segment. When the source and observation points are both located on the impedance boundary (i.e.,  $\rho = \rho' = a$ ), the modified Green's function  $G$  becomes

$$G = \frac{1}{i(\pi ka)^2} \int_C \frac{\exp(i\nu|\phi - \phi'|)}{[J'_\nu(ka) - iZ'J_\nu(ka)][H^{(2)}_\nu(ka) - iZ'H^{(2)}_\nu(ka)]} d\nu \quad (2)$$

where  $k$  is the free-space wavenumber, and the prime on the cylinder functions denotes the derivative with respect to the argument. The contour  $C$  and the singularities of the integrand in the complex  $\nu$ -plane are shown in Figure 1. In the lossless limit  $Z' = 0$ , this formulation reduces to the one in Reference 1.

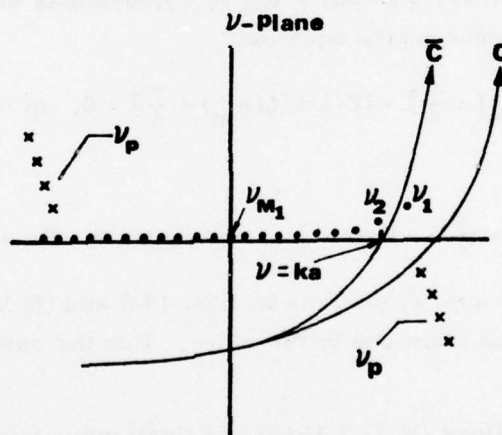


Fig. 1. Integration path and singularities in complex  $\nu$ -plane.  
 xx--zeros  $\nu_p$  of  $[H^{(2)}_\nu(ka) - iZ'H^{(2)}_\nu(ka)]$ ; ••-- zeros  
 $\nu_m$  of  $[J'_\nu(ka) - iZ'J_\nu(ka)]$ .

Anticipating subsequent contour deformation in the complex  $\nu$ -plane, we examine the location of the pole singularities  $\nu_m$  determined by the resonance equation

$$J'_\nu(ka) - iZ'J_\nu(ka) = 0, \quad m = 1, 2, \dots, M_1, \quad \text{Re } \nu_m > 0 \quad (3)$$

These poles in the integrand of Eq. (2) describe whispering gallery modes. For the special cases  $\arg Z' = \pm 90^\circ$ , the roots  $v_m$  are real. If for  $\arg Z' = -90^\circ$ , the magnitude  $|Z'|$  of the normalized impedance is greater than a certain value, one finds for the first root  $m=1$  that  $v_1 > ka$ . The corresponding field contribution then describes a surface wave that can exist also on a plane boundary. It is interesting to note that all of the whispering gallery modes represent fast waves when the boundary is perfectly conducting.<sup>1</sup> However, for  $Z' \neq 0$ , there may exist slow-wave type whispering gallery modes with  $\operatorname{Re} v_m > ka$ , which become surface waves when  $\operatorname{Im} v_m = 0$ , in addition. The resonance equation for the first  $\bar{M}$  whispering gallery modes bound most closely to the impedance surface is found to be

$$Ai'(t_m) + i\alpha Ai(t_m) = 0, \quad m = 1, 2, \dots, \bar{M}, \quad \alpha = Z' \left(\frac{ka}{2}\right)^{\frac{1}{3}} \quad (4a)$$

where the  $t_m$  are solutions of the differential equation

$$\frac{dt_m}{d\xi} = \frac{1}{t_m - \xi} Z, \quad \xi = -i\alpha \quad (4b)$$

In the other domain, where  $|v - ka| > 0$  ( $|v|^{\frac{1}{3}}$ ), one may employ the Debye approximations<sup>1</sup> to obtain the resonance equation

$$\cos w_m \cos \left[ \zeta(w_m) + \frac{\pi}{4} \right] - iZ' \sin \left[ \zeta(w_m) + \frac{\pi}{4} \right] = 0, \quad m = \bar{M} + 1, \dots, M_1 \quad (5)$$

where

$$\zeta(w) = ka \left[ \cos w - \left( \frac{\pi}{2} - w \right) \sin w \right], \quad v = ka \sin w, \quad \operatorname{Re} w > 0 \quad (5a)$$

It is found that the two approximations in Eqs. (4a) and (5) have an overlapping domain wherein one may switch from one to the other. For the case in Fig. 2,  $\bar{M} = 4$  has been found satisfactory.

## 2. Whispering Gallery (W.G.) Mode and Continuous Spectrum Representation

When the contour  $C$  in Fig. 1 is deformed into a contour extending along the imaginary  $v$ -axis, the Green's function can be represented as a sum of whispering gallery (W.G.) modes  $\bar{G}_m$  and a continuous spectrum  $R_{M_1}$ :

$$G = \sum_{m=1}^{M_1} \bar{G}_m + R_{M_1} \quad (6)$$

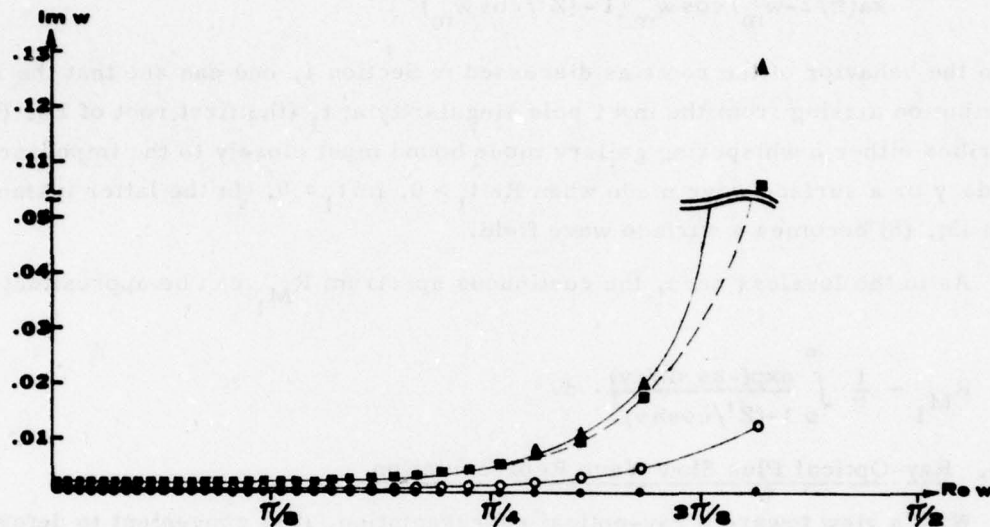


Fig. 2. Zeros of  $[J'_\nu(ka) - iZ'J_\nu(ka)]$  in complex  $w$ -plane. ●--zeros for  $Z'=0$  (perfect conductor); ○--zeros for  $Z'=0.05 \exp(-i30.5^\circ)$  (small loss); ■, ▲--zeros for  $Z'=0.22 \exp(-i30.5^\circ)$  (large loss). The first five roots ▲▲-- are obtained from (5b) and compared with ■■-- obtained from (7).

where

$$\bar{G}_m = \frac{J_\nu(ka) \exp[i\nu_m |\phi - \phi'| + i\pi/2]}{ka(\partial/\partial \nu) \{J'_\nu(ka) - iZ'J_\nu(ka)\}_{\nu=\nu_m}} \quad (7)$$

with the  $\nu_m$  satisfying the resonance equation in (3). Using the Fock type transition region approximations,<sup>1</sup> one finds

$$\bar{G}_m \sim \frac{\exp(iks + i\gamma t_m - i\pi/2)}{2(ka/2)^{1/3} (t_m + \alpha^2)} \quad , \quad m = 1 \dots \bar{M} \quad (8)$$

where

$$s = a|\phi - \phi'| \quad , \quad \gamma = \left(\frac{ka}{2}\right)^{1/3} \frac{s}{a} \quad (9)$$

Here,  $s$  denotes the path length between source and observation point along the boundary, and  $\gamma$  is an arc length parameter. For the remaining modes, Debye approximations and Eq. (5) may be used to obtain

$$\bar{G}_m \sim \frac{\exp(iks \sin w_m + i\pi/2)}{ka(\pi/2 - w_m) \cos w_m [1 - (Z'/\cos w_m)^2]}, \quad m = \bar{M} + 1 \dots M_1 \quad (10)$$

From the behavior of the roots as discussed in Section 1, one can see that the residue contribution arising from the  $m = 1$  pole singularity at  $t_1$  (the first root of Eq. (4)) describes either a whispering gallery mode bound most closely to the impedance boundary or a surface wave mode when  $\text{Re } t_1 > 0$ ,  $\text{Im } t_1 = 0$ . In the latter instance,  $\bar{G}_1$  in Eq. (8) becomes a surface wave field.

As in the lossless case, the continuous spectrum  $R_{M_1}$  can be approximated by<sup>1, 4</sup>

$$R_{M_1} \sim \frac{1}{\pi} \int_0^\infty \frac{\exp(-ks \sinh v)}{1 - (Z'/\cosh v)^2} dv \quad (11)$$

### 3. Ray-Optical Plus Slow Wave Representation

With a view toward a ray-optical representation, it is convenient to deform the contour  $C$  in Fig. 1 into  $\bar{C}$ , whence  $G$  becomes

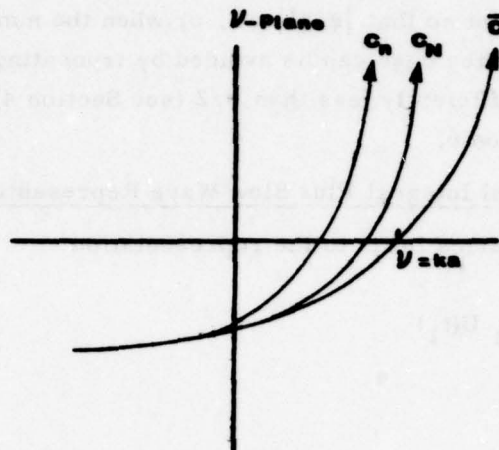
$$G = \bar{G} + \bar{G}_1 U(t_1) \quad (12)$$

where  $\bar{G}$  is the same as  $G$  except for the replacement of  $C$  by  $\bar{C}$  in Eq. (2), and  $\bar{G}_1$  is the contribution (if any) from the pole singularity at  $t_1$  located between  $C$  and  $\bar{C}$ . The unit step function  $U(t_1)$  contributes only when the  $t_1$  pole lies between the two contours. The extraction of  $\bar{G}_1$ , which may be either a slow wave  $W$ ,  $G$ , mode or a surface wave, is desirable since slow wave fields do not exhibit ray-optical properties. Thereafter, the Bessel function denominator in Eq. (2) is expanded into a "traveling wave" series, Debye approximations are employed, and the resulting integrals are evaluated along steepest descent paths  $C_n$  passing through saddle points  $v_{sn}$  in the complex  $v$ -plane (Fig. 3). The result is

$$\bar{G} = \sum_{n=0}^{\infty} \bar{G}_n \quad (13)$$

where

$$\bar{G}_n \sim e^{i\pi/4} \sqrt{\frac{2}{\pi k}} (-i)^n \frac{(\Gamma_n)^n \cos^2 w_{sn} \exp(ikD_n)}{(\cos w_{sn} + Z')^2 \sqrt{D_n}} \quad (14)$$

Fig. 3. Steepest descent paths in complex  $v$ -plane.

$$D_n = 2(n+1) a \sin [|\phi - \phi'|/2(n+1)] , \quad w_{sn} = \frac{\pi}{2} - \frac{|\phi - \phi'|}{2(n+1)} \quad (14a)$$

and  $\Gamma_n$  is the boundary reflection coefficient for ray species  $n$ ,

$$\Gamma_n = \frac{\cos w_{sn} - Z'}{\cos w_{sn} + Z'} \quad (14b)$$

The expression in Eq. (14) corresponds to a geometric-optical ray field that has undergone  $n$  reflections at the impedance boundary (see Fig. 4). The sum in Eq. (13) can

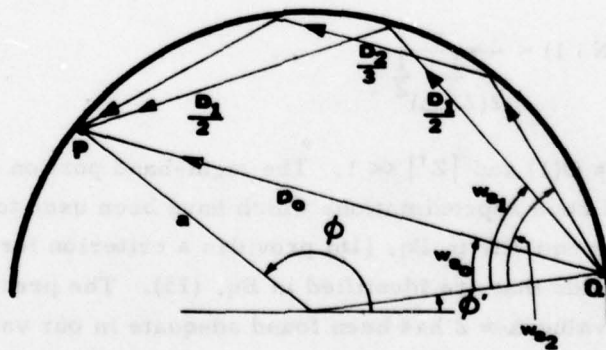


Fig. 4. Direct and multiply reflected rays for a circular boundary. Geometrical quantities  $D_0$ ,  $D_1$ ,  $D_2$ ,  $w_{s0}$ ,  $w_{s1}$  and  $w_{s2}$  are also shown.

actually not be extended to  $n \rightarrow \infty$  since the Debye approximations used to obtain Eq. (14) become invalid as  $w_{sn} \rightarrow \pi/2$ . This situation arises either when the observation point approaches the source point so that  $|\phi - \phi'| \rightarrow 0$ , or when the number of rays increases, with  $|\phi - \phi'|$  fixed. The latter case can be avoided by truncating the number of rays at some  $n = N$  such that  $w_{sN}$  is sufficiently less than  $\pi/2$  (see Section 4) while the former case will be discussed in Section 6.

#### 4. Ray Plus Canonical Integral Plus Slow Wave Representation

Truncating the ray series leads to the representation

$$G = \sum_{n=0}^N \tilde{G}_n + R_N + \bar{G}_1 U(t_1) \quad (15)$$

where

$$R_N = \frac{1}{i(\pi ka)^2} \int \frac{(-r)^{N+1} \exp(i\nu|\phi - \phi'|)}{\bar{C} [H_{\nu}^{(2)}(ka) - iZ'H_{\nu}^{(2)}(ka)] [J'_{\nu}(ka) - iZ'J_{\nu}(ka)]} d\nu \quad (15a)$$

$\tilde{G}_n$ , given in Eq. (17), again describes a geometrical ray field having undergone  $n$  reflections. The remainder term  $R_N$  accounts for ray fields reflected more than  $N$  times.

One way of dealing with Eq. (15a) is to investigate how fast the integrand decays in the neighborhood of the point  $\nu = ka$ ; rapid decay permits effective truncation of the integration path and the use of further approximations. Proceeding as in the  $Z' = 0$  case,<sup>1</sup> one may show that the main contribution to the integral arises from the neighborhood of  $\nu = ka$  provided that the arc length parameter  $\gamma$  defined in Eq. (9) and the number of rays  $(N+1)$  satisfy the left-hand portion of the inequality

$$\frac{\gamma}{2(2^{\frac{1}{3}} \Delta)^{\frac{1}{2}}} - 1 < (N+1) < \frac{\gamma}{2(2^{\frac{1}{3}} \Delta)^{\frac{1}{2}}} \quad (16)$$

where  $\Delta = \Delta(ka, Z') = O(1)$  and  $|Z'| \ll 1$ . The right-hand portion of the inequality validates use of the Debye approximations which have been used to derive  $\tilde{G}_n$  in Eq. (14). The complete inequality in Eq. (16) provides a criterion for the number of geometric optical ray fields that are identified in Eq. (15). The precise choice of  $\Delta$  is not too important; a value  $\Delta \approx 2$  has been found adequate in our various numerical comparisons<sup>1</sup> (Section C).

Since the contributing portion of the integrand in Eq. (15a) is localized near  $\nu = ka$ , one may substitute the Fock type asymptotic approximations to obtain

$$R_N = \frac{(-1)^{N+1} \exp(iks)}{2i\pi^2 ka} \left(\frac{ka}{2}\right)^{\frac{2}{3}} I_N(\gamma, \alpha) \quad \alpha = Z' \left(\frac{ka}{2}\right)^{\frac{1}{3}} \quad (17)$$

Here,  $I_N(\gamma, \alpha)$  is the canonical integral

$$I_N(\gamma, \alpha) = \int_{\infty e^{i4\pi/3}}^{\infty e^{i\pi/3}} \frac{[\{w_1'(t) + i\alpha w_1(t)\} / \{w_2'(t) + i\alpha w_2(t)\}]^{N+1}}{[w_2'(t) + i\alpha w_2(t)] [A_1'(t) + i\alpha A_1(t)]} e^{i\gamma t} dt \quad (18)$$

which has been evaluated for relevant ranges of  $\gamma$  and three values of  $\alpha$ . The results are shown in Fig. 11 and have been tabulated.<sup>5</sup> The canonical integral for a related boundary value problem corresponding to a different integrand and to  $\alpha = 0$  has been tabulated by Babich and Buldyrev.<sup>6</sup>

#### 5. Ray Plus Whispering Gallery (W.G.) Mode Plus Slow Wave Representation

An alternative way of treating Eq. (15a) is to deform the contour  $\bar{C}$  into  $C_N$  in Fig. 3 and thereby representing  $R_N$  as a sum of  $[M - U(t_1)]$  W.G. modes whose poles lie between  $\bar{C}$  and  $C_N$ , plus some other remainder term  $R_{MN}$ . One finds

$$R_N = \sum_{m=\bar{m}}^M \bar{C}_m + R_{MN}, \quad \bar{m} = 1 + U(t_1) \quad (19)$$

where

$$R_{MN} = -\frac{1}{2} \bar{C}_N \left(1 - \frac{Z'}{\cos w_{sN}}\right) \quad (19a)$$

Thus

$$G = \sum_{n=0}^N \bar{C}_n + \sum_{m=\bar{m}} \bar{C}_m - \frac{1}{2} \bar{C}_N \left(1 - \frac{Z'}{\cos w_{sN}}\right) \quad (20)$$

#### 6. Near Field Form

As observed at the end of Section 3 (see also Eq. (16)), any Green's function representation that includes geometric optical terms fails in the near field where the arc length parameter  $\gamma$  in Eq. (9) is less than  $2(2^{1/3} \Delta)^{1/2}$ . For this case, one may find the near field representation:

$$G = \frac{e^{iks + i\pi/4}}{\sqrt{2\pi ks}} \left\{ 1 + \sum_{j=1}^{21} b_j \gamma^{j/2} + O(\gamma^{11}) \right\}, \quad |\alpha| \leq 1 \quad (21)$$

where

$$b_j = \begin{cases} \sqrt{\pi} a_j e^{i(\frac{\pi}{4}j)} / (j-1) & j \text{ odd} \\ 2^{j/2} a_j e^{i(\frac{\pi}{4}j)} / \sum_{m=1}^{j/2} (2m-1), & j \text{ even} \end{cases}$$

The coefficients  $a_j$  have been given elsewhere.<sup>7</sup> The restriction  $|\alpha| \leq 1$  in Eq. (21) arises from the condition  $|\alpha Ai(t)/Ai'(t)| < 1$ . In view of the definition of  $\alpha$  in Eq. (4a), the solution Eq. (21) can, therefore, not be applied to the infinite plane limit  $a \rightarrow \infty$  if  $Z'$  differs from zero.

### C. Numerical Results

Extensive numerical computations have been performed for  $ka = 100$  and for various surface impedance values to show how the amplitude and phase of the surface impedance affect the field, and to check the accuracy and range of validity of the various formulations in Section B.

Concerning the propagation constants for the W. G. modes, we have obtained solutions of the differential equation (5b) by the Runge-Kutta method. The results are shown in Figure 5. Here  $t_1$ ,  $t_2$  and  $t_3$  are the first three roots of  $Ai'(t_m) = 0$  (i.e.,  $\alpha = 0$  in Eq. (5a) and  $\tilde{t}_1$ ,  $\tilde{t}_2$  and  $\tilde{t}_3$  are those of  $Ai(\tilde{t}_m) = 0$  (i.e.,  $\alpha = \infty$  in Eq. (5a)). The solid lines and the dotted lines are equi-phase and equi-amplitude curves, respectively. The arrows on the equi-phase curves indicate the directions along the root loci with increasing amplitude  $|\xi|$ . It is interesting to note in Fig. 5 that the first zero  $t_1$  crosses the imaginary axis for some values of  $\xi$ , thereby, furnishing the slow wave type of whispering gallery mode. In particular, when  $\text{Re } t_1 > 0$ ,  $\text{Im } t_1 = 0$ , the contribution from  $t_1$  describes a surface wave mode. Except for  $\arg Z' = \pm 90^\circ$ , the imaginary parts of the whispering gallery mode poles become large as the amplitude of  $Z'$  increases within the ranges under consideration (i.e.,  $|Z'| < 0.6$ ). It should be noted in Figs. 2 and 5 that, for given  $Z'$  and  $\arg Z' \neq \pm 90^\circ$ , the eigenvalues for modes bound close to the surface have large imaginary parts compared to those located far from the surface, thereby de-emphasizing the importance of the most closely bound whispering gallery modes. This is to be expected on physical grounds since the field of the tightly bound modes is in close proximity to the lossy guiding surface. It can also be seen from Fig. 2 that the two approximations given in Eqs. (5a) and (7) have an overlapping region wherein one can switch from one to the other.  $\bar{M} = 4$  has been found satisfactory for our evaluation of the field.

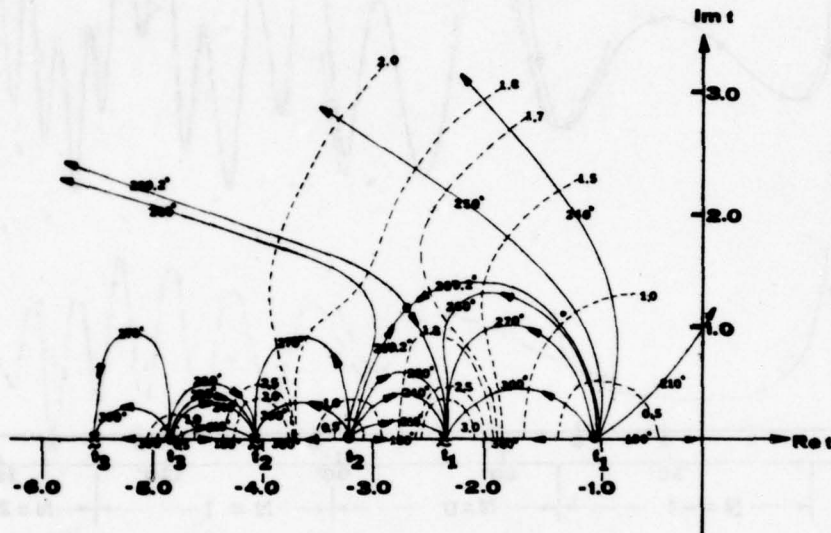


Fig. 5. Roots of  $[Ai'(t_m) + i\alpha Ai(t_m)]$  as obtained from (5b). The numbers on the solid and dashed curves denote the phase and amplitude of  $\xi = -i\alpha$ , respectively.

In Figs. 6 and 7, we have compared the various formulations in Section B for a purely inductive impedance case ( $|Z'| = 0.22$  and  $\arg Z' = -90^\circ$ ) to check accuracy and range of validity. For this case, the surface wave mode is excited since the first root of Eq. (5a) or Eq. (5b) is located on the positive real axis in the  $t$ -plane. The ray plus canonical integral plus slow wave representation in Eq. (19) is seen to provide an excellent approximation provided that the number of rays ( $N+1$ ) is chosen according to the criterion in Eq. (20). Because of the overlap of the curves for various  $N$ , switching from one formulation to the other can be performed smoothly. The ray plus canonical integral contribution curves are shown separately to exhibit the effect of the surface wave on the total field.

In Fig. 7, the near field form in Eq. (30) and the ray plus whispering gallery mode plus slow wave representation in Eq. (27) are compared with the whispering gallery mode (including surface wave) plus continuous spectrum representation in Eq. (8). It is interesting to observe that for large enough  $\gamma$ , the total field can be accurately represented just by the surface wave ( $M=1$ ) and by rays provided that the appropriate number of reflected rays is included as  $\gamma$  or  $|\phi - \phi'|$  increases.

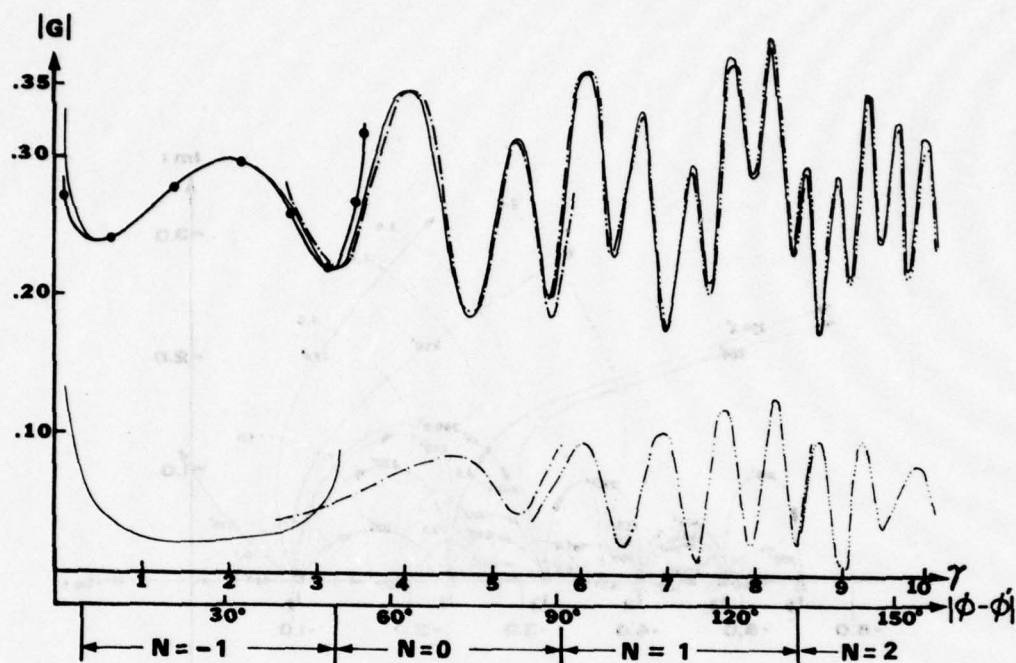


Fig. 6. Ray plus canonical integral plus slow wave representation in Eq. (19), with  $Z' = 0.22 \exp(-i90^\circ)$ . Both the  $\gamma$  and  $|\phi - \phi'|$  coordinates are indicated. Also shown are the ranges in  $\gamma$  and  $|\phi - \phi'|$  corresponding to  $N = -1$  (—•—•—) (no geometric-optical ray);  $N = 0$  (—•—) (1 ray);  $N = 1$  (—••—) (2 rays);  $N = 2$  (—•••—) (3 rays). The solid curve is calculated from Eq. (8) and serves as the reference solution. The heavy curves include the surface wave in Eq. (19) ( $U(t_1) = 1$  for this case), while the light curves have been obtained from Eq. (19) with the surface wave omitted.

Curves for a high loss case are depicted in Fig. 8, based on the ray plus canonical integral representation. Having in mind an application to ground wave propagation, the high loss impedance  $Z' = 0.22 \exp(-i30.5^\circ)$  has been obtained by assuming a wet ground surface with conductivity  $\sigma_e = 10^{-2} \text{ mho m}^{-1}$  and dielectric constant  $\epsilon_e = 10 \epsilon_0$ , subject to a wave frequency  $f = 10 \text{ MHz}$ . Also shown in Fig. 8 is the near field form for small  $\gamma$ . It should be noted that  $G_1$  in Eq. (19) does not contribute since no pole singularity is located between the contours  $C$  and  $\bar{C}$  in Fig. 1 (see also Section B-1). All of the modes here are fast waves as seen from Fig. 2. Because of the dissipative attenuation of the modes bound close to the surface, the canonical integral contribution is found to be negligible for  $\gamma > 4$ , thereby establishing the adequacy of the ray-optical contribution alone for describing the field sufficiently far from the source point.

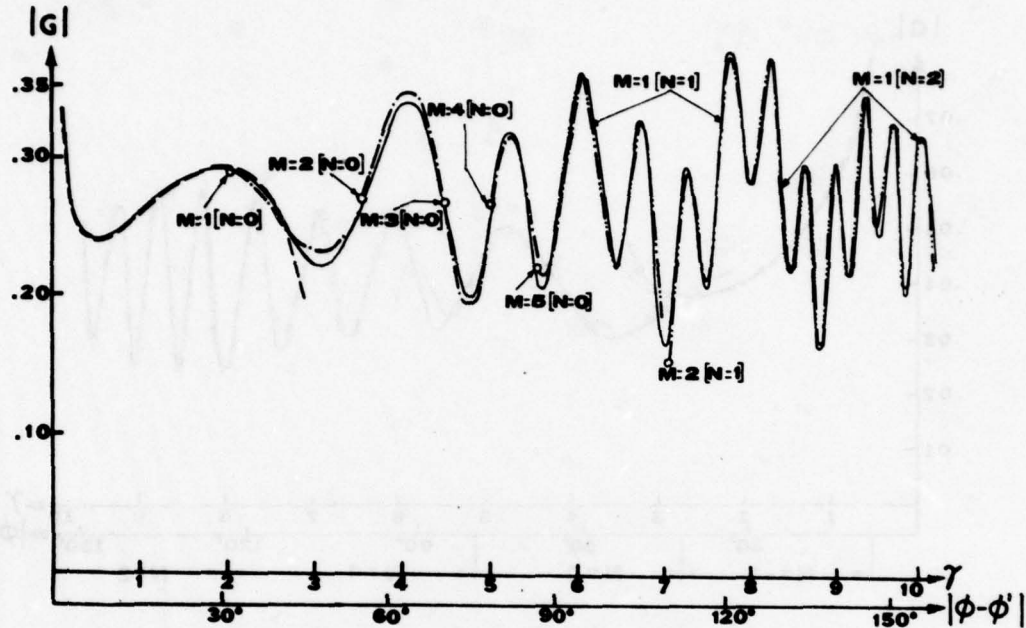


Fig. 7. Ray plus W. G. mode plus slow wave representation in Eq. (27), and near field form in Eq. (30) (---), with  $Z' = 0.22 \exp(-i90^\circ)$ . The reference solution (solid curve) is calculated from Eq. (8). The numbers indicated along the curves should be read as follows: for example,  $M = 2 [N=1]$  denotes the range wherein the direct ray and singly reflected rays plus 2 modes (1 W. G. mode and 1 surface wave mode) are applicable, while the circles indicate the starting point of the relevant intervals. The  $|\phi - \phi'|$  coordinate is not shown since it has been depicted in Fig. 6.

The solid curves in Figs. 6, 7 and 8 are obtained from Eq. (8) with  $M_1 = 32$  ( $ka = 100$ ). This form of the solution is taken as a reference for all  $\gamma$ . It was noted in the study of the perfect conductor case<sup>1</sup> that the numerical values derived from the W. G. mode plus continuous spectrum representation are extremely sensitive to the exact number of modes required. Unless all 32 modes (for  $ka = 100$ ) are included, the perfect conductor was found to deviate appreciably from the correct shape. The same remark applies to the finite surface impedance case in Figs. 6, 7 and 8 since the higher order modes have small attenuation coefficients (see Fig. 2) even though the losses may be large. Because of this feature, it is preferable to use one of the other representations for field calculation.

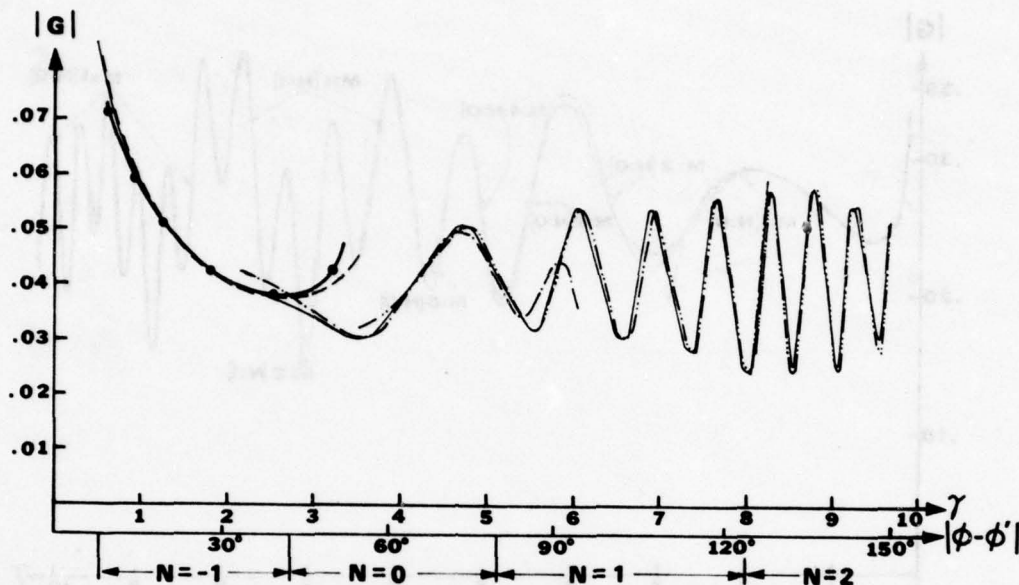


Fig. 8. Ray plus canonical integral representation in Eq. (19), and near field form (--) in Eq. (30), with  $Z' = 0.22 \exp(-i30.5^\circ)$ . Note that  $U(t_1) = 0$  in this example (see Fig. 2). The reference solution (solid curve) is calculated from Eq. (8).  
 —•—•—  $N = -1$ ; —•—  $N = 0$ ; —•—  $N = 1$ ; and —•—  $N = 2$ .

In Fig. 9, we have shown how increasing the amplitude of the normalized impedance  $Z'$  affects the surface field; here  $\arg Z' = -45^\circ$ . Curves are depicted based on the ray plus canonical integral representation, and on the near field form for small  $\gamma$ . In our computations,  $|Z'| = 0.05$  has been chosen as a sample of a small impedance, and  $|Z'| = 0.22$  as a sample of a large impedance. When  $\arg Z' = -90^\circ$ , corresponding to a purely inductive impedance boundary which supports a slow wave, the field amplitude is found to increase as  $|Z'|$  increases while the converse is true when the boundary impedance is purely capacitive (see Reference 7). This behavior may be attributed to the greater and lesser confinement, respectively, of the lowest order modal field in the two cases. As the imaginary part of roots increases (see Fig. 5), the field transmitted to an observation point tends to decrease (Figure 9). Because of the absorption on the surface, the interference effects between rays are also weakened, as seen from the decreasing magnitude of oscillation in the curves in Figure 9.

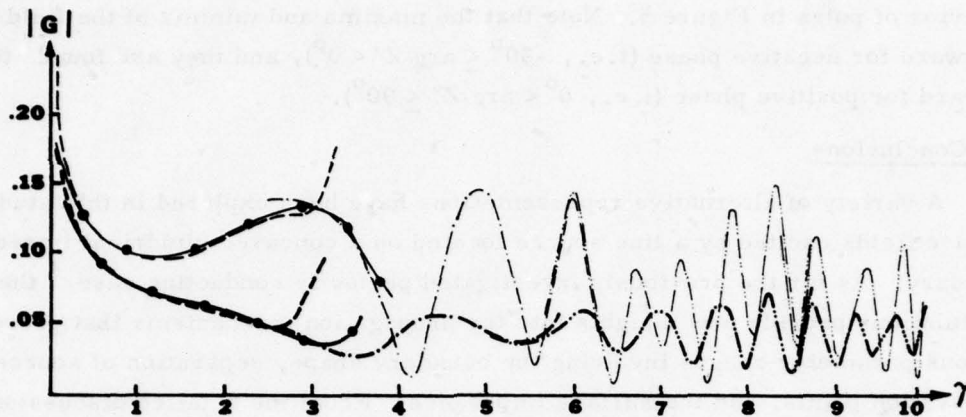


Fig. 9. Influence of the amplitude of the surface impedance  $Z'$ , calculated from Eqs. (19) and (30) (---), for  $\arg Z' = -45^\circ$ . The heavy curves represent a large impedance ( $|Z'| = 0.22$ ) and the light curves a small impedance ( $|Z'| = 0.05$ ). Here,  $U(t_1) = 0$ .  
 —•—•—  $N = -1$ ; ---  $N = 0$ ; -.-.  $N = 1$ ; and  
 ....  $N = 2$ .

The curves in Fig. 10 show the effect of the phase of  $Z'$  on the field when the amplitude remains constant at  $|Z'| \cong 0.05$ . One observes that the field tends to increase as the phase of  $Z'$  deviates from zero; this could have been predicted from the

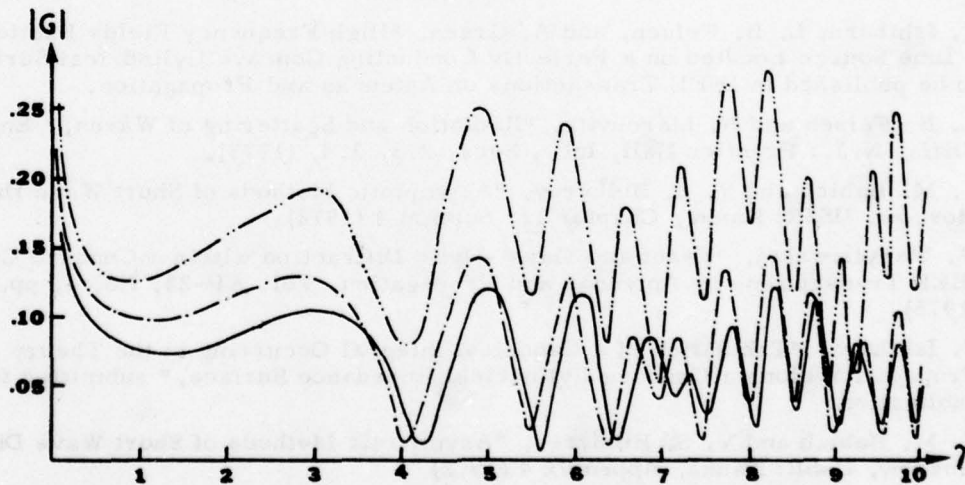


Fig. 10. Influence of the phase of the surface impedance  $Z'$ , calculated from Eq. (19), with  $|Z'| = 0.05$ .  
 —  $\arg Z' = 0^\circ$ ; ---  $\arg Z' = -45^\circ$ ; -.-.  $\arg Z' = -90^\circ$ .

behavior of poles in Figure 5. Note that the maxima and minima of the field shift backward for negative phase (i.e.,  $-90^\circ \leq \arg Z' < 0^\circ$ ), and they are found<sup>7</sup> to shift forward for positive phase (i.e.,  $0^\circ < \arg Z' \leq 90^\circ$ ).

#### D. Conclusions

A variety of alternative representations have been explored in this study of the surface fields excited by a line source located on a concave cylindrical impedance boundary. As for the previously investigated perfectly conducting case,<sup>1</sup> these different formulations provide new insights into the propagation mechanisms that prevail for various parameter ranges involving the boundary shape, separation of source and observation points, and the surface impedance. From the detailed discussion and interpretation of the numerical results in Section C emerges the role played by the geometric optical field and, for sufficiently inductive boundaries, by the surface wave. The presence of the surface wave leads to an enhancement of the field over that observed on a perfect conductor. When losses are appreciable, attenuation of mode fields bound close to the boundary establishes the geometric optical field as the dominant and adequate constituent. The versatility of representations involving the canonical integral has again been confirmed, and tabulations have been provided for several values of surface impedance.

Army Research Office  
DAHC-04-75-G-0152

T. Ishihara and L. B. Felsen

#### REFERENCES

1. T. Ishihara, L. B. Felsen, and A. Green, "High Frequency Fields Excited by a Line Source Located on a Perfectly Conducting Concave Cylindrical Surface," to be published in IEEE Transactions on Antennas and Propagation.
2. L. B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," Englewood Cliffs, N.J.: Prentice Hall, Inc., Secs. 2.3, 3.4, (1973).
3. V. M. Babich and V. S. Buldyrev, "Asymptotic Methods of Short Wave Diffraction," Moscow, USSR: Nauka, Chapter 11, Section 4 (1972).
4. W. Wasylkiwskyj, "Exact and Quasi-Optic Diffraction within a Concave Cylinder," IEEE Transactions on Antennas and Propagation, Vol. AP-23, No. 4, pp. 480-492 (1975).
5. T. Ishihara, "Tabulation of a Canonical Integral Occurring in the Theory of Wave Propagation along a Concave Cylindrical Impedance Surface," submitted for publication.
6. V. M. Babich and V. S. Buldyrev, "Asymptotic Methods of Short Wave Diffraction," Moscow, USSR: Nauka, Appendix 4 (1972).
7. T. Ishihara, "High Frequency Behavior of Source Excited Concave Surfaces," Ph.D. Dissertation (Electrophysics), Polytech. Inst. of New York, (June 1978).
8. G. Hasserjian and A. Ishimaru, "Currents Induced on the Surface of a Conducting Circular Cylinder by Slot," J. Res. Nat. Bur. Stand., Vol. 66D, No. 3, pp. 335-365 (1972).

## HIGH-FREQUENCY SURFACE FIELDS EXCITED BY A POINT SOURCE ON A CONCAVE PERFECTLY CONDUCTING CYLINDRICAL BOUNDARY

L. B. Felsen and T. Ishihara

A. Introduction and Summary

As noted previously, when a high-frequency radiator is located on or near a concave surface, the fields on or near the surface cannot be calculated by conventional geometrical optics, although the observation point is visible from the source. This circumstance has led to a detailed study of alternative formulations for evaluation of the surface field. The interior of a circular cylinder has served as a prototype because of the availability of rigorous field solutions for this configuration. A detailed study of the two-dimensional axially independent problem<sup>1,2</sup> has revealed the utility of field representations that incorporate in various combinations geometric optical contributions, whispering gallery modes, continuous spectra, and canonical wave species analogous to those expressed by the Fock integral for convex shapes. This previous investigation has provided not only a quantitative means for determining the field but also new insight into the propagation characteristics.

In the present study, the analysis is generalized in Section B to three dimensions by considering fields excited by an axial magnetic dipole source instead of a line source. The perfectly conducting circular cylindrical configuration is retained for the reasons stated above. Because of a simple Fourier integral relation that connects two-dimensional and three-dimensional Green's functions in axially uniform cylindrical regions, the previous results<sup>1</sup> can be utilized directly in the analysis. Therefore, the presentation here is limited to providing only the major steps and the results since details may be obtained from Reference 1. Emphasis is placed on the derivation of tractable, numerically accurate field expressions (Section C) that incorporate "invariant" parameters. This feature furnishes not only a physical interpretation of the propagation mechanism but also facilitates subsequent generalization to non-circular shapes. The utility and accuracy of these expressions is assessed in Section D by extensive numerical calculations which also highlight limitations that inhibit easy extension of certain two-dimensional formulations to three dimensions.

A limitation of the results presented here is the inability to accommodate observation points displaced axially or nearly axially from the source. As for the convex surface, the field behavior in the paraxial region cannot be inferred from generalization of axially independent solutions.

### B. Analytical Formulation

The analytical formulation of the three-dimensional problem from a knowledge of the two-dimensional one is straightforward. The time-harmonic three-dimensional Green's function  $G(\underline{r}, \underline{r}', k)$ , where  $\underline{r} = (\rho, z)$  is the observation point,  $\underline{r}' = (\rho, z')$  is the source point, and  $k$  is the free-space wavenumber, can be obtained from the two-dimensional  $z$  independent Green's function  $\bar{G}(\rho, \rho', k)$  by the following operation:<sup>3</sup>

$$G(\underline{r}, \underline{r}', k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \bar{G}(\rho, \rho', k_t) e^{ik_z |z-z'|} dk_z \quad (1)$$

where

$$k_t = \sqrt{k^2 - k_z^2}, \quad \text{Im } k_t \geq 0 \quad (1a)$$

The dependence on  $k$  in  $G$  will henceforth be omitted in the notation. A time dependence  $\exp(-i\omega t)$  is implied. The Green's function  $G$  is assumed to have a vanishing normal derivative on the cylindrical boundary, and thus represents a scalar potential from which the electromagnetic field of an axially oriented magnetic dipole element can be derived.

When the transformation in Eq. (1) is applied to asymptotic or other approximate forms of  $\bar{G}$ , care must be taken to ensure that the replacement in  $\bar{G}$  or  $k$  by  $k_t$ , with  $-\infty < k_z < \infty$ , does not invalidate assumptions made in the approximate evaluation of  $\bar{G}$ . In Ref. 1, it was assumed that  $ka \gg 1$ , where  $a$  is the radius of curvature of the cylindrical surface. Thus, the integration path in Eq. (1) must be deformed away from the branch point at  $k_z = k$ . Moreover, the major contribution to the integral in Eq. (1) must come from  $k_z$  values that retain  $|k_t a|$  large. Since  $k_t$  represents the wavenumber transverse to  $z$ , this implies that all relevant wave constituents must have a large transverse component, i.e., observation points displaced axially or almost axially from the source point are excluded. Furthermore, if complex values of  $k_t$  should become relevant, the appropriate analytic continuation from  $k$  to  $k_t$  must be performed. Subject to these considerations, most of the two-dimensional results can be readily transformed via Eq. (1) into the three-dimensional domain.

A variety of different field representations, each with its own physical interpretation and utility, can be extracted from Eq. (1). They are obtained from the corresponding representations in Ref. 1 upon eliminating the  $k_z$ -integral, when possible, by saddle point evaluation with the aid of the transformation from the  $k_z$ -plane to the  $\theta$ -plane via  $k_z = k \sin \theta$ , with  $k_t = k \cos \theta$ . Any two-dimensional field constituent that is

of the form

$$\bar{G} = B(\rho, \rho', k) \exp(ikD) \quad (2)$$

where  $B$  is a slowly varying amplitude and  $D$  is a length parameter, gives rise to a saddle point  $k_{zs}$  in Eq. (1) located at

$$k_{zs} = k \sin \theta_s, \quad \theta_s = \tan^{-1} \left( \frac{|z-z'|}{D} \right) \quad (2a)$$

with  $k_{ts} = k \cos \theta_s$ . In the  $\theta$ -plane, the saddle point is located at  $\theta_s$ . Because of the restrictions noted earlier, the range  $\theta_s \rightarrow \pi/2$  must be excluded. When the integral in the  $\theta$ -plane is then evaluated by the conventional saddle point formula,<sup>4</sup> the resulting three-dimensional field has the form:

$$G(\underline{r}, \underline{r}') \sim \sqrt{\frac{k \cos^2 \theta_s}{2\pi L}} B(\rho, \rho', k \cos \theta_s) e^{ikL - i\pi/4} \quad (3)$$

where

$$L = |z-z'| \sin \theta_s + D \cos \theta_s \quad (3a)$$

### C. Alternative Field Representations

#### 1. Basic Representation

The basic integral representation is given by Eq. (1) with<sup>1</sup>

$$\bar{G}(\rho, \rho', k_t) = \frac{1}{i(\pi k_t a)^2} \int_{\bar{C}} \frac{\exp[i\nu|\phi - \phi'|]}{H'_\nu(2)(k_t a) J'_\nu(k_t a)} d\nu \quad (4)$$

where the contour  $\bar{C}$  and the singularities of the integrand in the complex  $\nu$ -plane are shown in Fig. 1, and the prime on the Bessel and Hankel functions denotes the derivative with respect to the argument. The source and observation points lie on the large circular cylindrical boundary of radius  $a$  whereon the normal derivative of  $G$  is assumed to vanish. Note that the formulation in Eqs. (1) and (4) contains the essential information for the fields along a circular concave segment. All effects that would represent behavior attributable to a closed cylindrical cavity have been removed.<sup>1</sup> In particular, the azimuthal periodicity has been eliminated by formulating the field problem in an infinite angular space  $-\infty < \phi < \infty$ . This formulation, equivalent to placing "perfect absorbers for angularly propagating waves" along two radial planes, introduces spurious scattering from the origin, which has also been extracted.



where

$$R_N = \frac{(-1)^{N+2} e^{ikR + i\pi/4}}{2\pi^2 a \sqrt{2\pi kR}} \left( \frac{ka \cos \theta_s}{2} \right)^{2/3} I_N(\beta) \quad (9a)$$

$$I_N(\beta) = \int_{-\infty e^{i4\pi/3}}^{\infty e^{i\pi/3}} \frac{e^{i\beta t}}{A_1'(t) W_2'(t)} \left[ \frac{W_1'(t)}{W_2'(t)} \right]^{N+1} dt \quad (9b)$$

and

$$\beta = \left( \frac{ka \cos \theta_s}{2} \right)^{1/3} |\phi - \phi'| = \left( \frac{ka \cos \theta_s}{2} \right)^{1/3} \frac{s}{a} \quad (10)$$

$$R = |z - z'| \sin \theta_s + s \cos \theta_s \quad (11)$$

The number of rays  $(N+1)$  and the arc length parameter  $\beta$  are related by the criterion

$$\frac{\beta}{2(2^{1/3} \Delta)^{1/2}} - 1 < (N+1) < \frac{\beta}{2(2^{1/3} \Delta)^{1/2}} + 1, \quad \Delta = \Delta(ka) = O(1) \quad (12)$$

The geometrical quantities  $R$ , the distance between  $Q$  and  $P$  along the geodesic on the cylinder surface, and  $\theta_s$ , the geodesic departure angle, are shown in Fig. 2. Except for the replacement of  $\gamma$  by  $\beta$ ,  $I_N(\beta)$  is the same as  $I_N(\gamma)$  and has been tabulated elsewhere<sup>2,5</sup> for selected ranges of  $\beta$  and  $N$ .

#### 4. Ray plus Whispering Gallery (W.G.) Mode Representation

It may be shown that the remainder integral in Eq. (9) can be simplified so that one obtains instead of Equation (8):

$$G = \sum_{n=0}^N \tilde{G}_n + \sum_{m=1}^M G_m - \frac{1}{2} \tilde{G}_N \quad (13)$$

The  $m$ -th three-dimensional W. G. mode  $G_m$  is given by

$$G_m = \frac{e^{ikR + i\pi/4}}{\sqrt{2\pi a^2 kR}} \left( \frac{ka \cos \theta_s}{2} \right)^{2/3} \frac{e^{-i\beta \sigma_m}}{\sigma_m} \quad (14)$$

where  $\sigma_m$  are determined by

$$Ai'(-\sigma_m) = 0, \quad m = 1, 2, \dots, M \quad (14a)$$

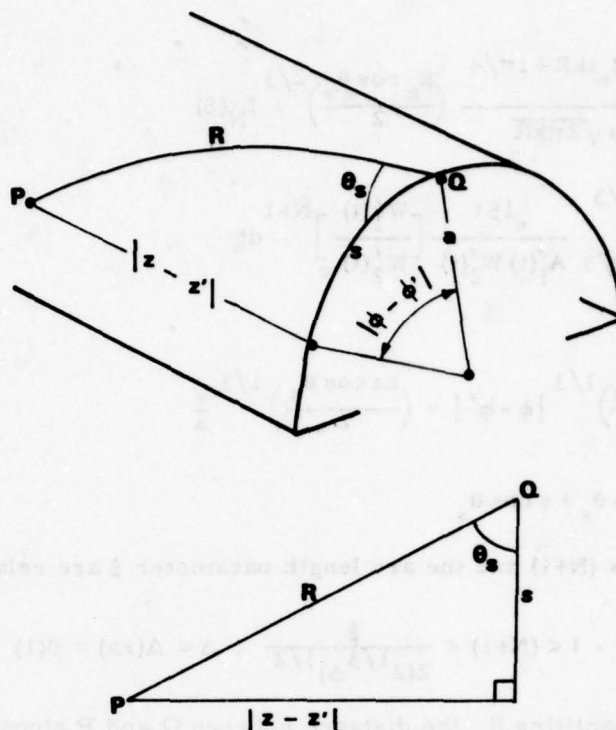


Fig. 2. Geometrical quantities pertaining to whispering gallery mode.  $R = R(\theta_s)$  is the geodesic along the surface between the source and observation points.

The number of rays ( $N+1$ ) is determined by the inequality Equation (12). Because of Eq. (12), larger values of  $\beta$  are required for inclusion of higher order reflected rays. For given  $N$ , as  $\beta$  increases in its applicable range, a new W.G. mode must be included whenever the  $N$ -times reflected ray touches the caustic of that mode. The process for proper selection of  $N$  and  $M$  is schematized in Fig. 3, where each concentric cylinder depicts the modal caustic of a three dimensional whispering gallery mode. For example, within the ranges of  $P_1$ ,  $P_2$  and  $P_3$ , one may choose  $N=0$  ( $M=1$ ),  $N=0$  ( $M=3$ ) and  $N=1$  ( $M=1$ ), respectively. It should be noted that in the simpler two-dimensional case,<sup>1</sup> we were able to describe the field by a single (direct) ray plus W.G. modes for  $\gamma \geq 2(2^{1/3}\Delta)^{1/2} \approx 3.1$  since it was possible to adequately represent the required number of modes, even those of higher order. However, in the three-dimensional case, use of the Fock asymptotic approximations restricts the applicability of the formulas to the lower order modes. This aspect is discussed further in Section D.

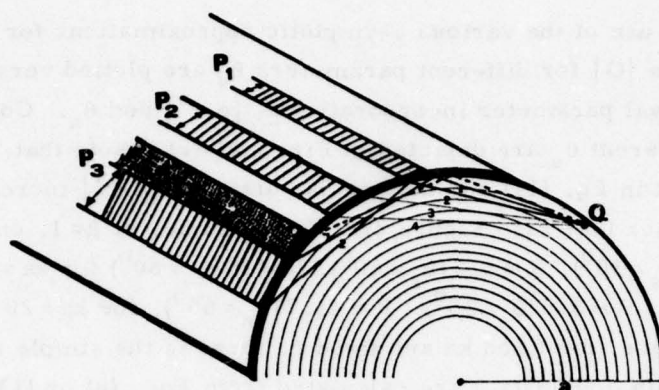


Fig. 3. Geometrical interpretation of ray plus whispering gallery mode representation.

#### 5. Near Field and Infinite Plane Limits

None of the above representations applies in the limiting case  $\beta \rightarrow 0$  in Eq. (10), which arises either when  $s$  and  $|z - z'| \rightarrow 0$  with  $\underline{a}$  and  $\theta_s$  fixed, or when  $a \rightarrow \infty$  with  $s$  and  $\theta_s$  fixed. For this case, one may show that<sup>1</sup>

$$G \sim \frac{e^{ikR}}{2\pi R} \left\{ 1 + \sum_{j=1}^{10} b_j \beta^{(3j/2)} + O(\beta^{33/2}) \right\} \quad (15)$$

#### 6. Whispering Gallery Mode plus Continuous Spectrum Representation

In the two-dimensional case, it was feasible to derive a field representation in terms of the totality of W. G. modes plus a continuous spectrum that accounts primarily for near field effects.<sup>1</sup> This formulation served as an accurate, though not necessarily convenient, reference solution since it was possible to render both of these fields constituents numerically tractable. However, the additional integration in Eq. (1) required for the three-dimensional case introduces complications that make the numerical treatment impractical and inaccurate. For this reason, we have omitted this representation from the present study.

#### D. Numerical Results

Numerical computations have been performed for a circular cylindrical boundary with  $ka = 100, 50, 20$  and initial geodesic departure angles  $\theta = \theta^0, 30^\circ, 60^\circ$  and  $80^\circ$  ( $ka = 100$ ). Because of the restriction that the main contributions to the basic integral in Eq. (1) must arise from  $k_z$  values that retain  $|k_t a|$  large, we have chosen the largest geodesic departure angle as  $\theta_s = 80^\circ$  for  $ka = 100$  and as  $\theta_s = 60^\circ$  for  $ka = 50$  and

20. This validates use of the various asymptotic approximations for  $\bar{G}$ . In each figure, the field magnitudes  $|G|$  for different parameters  $\theta_s$  are plotted versus  $\beta$  in Eq. (10), which is the universal parameter incorporating  $\underline{a}$ ,  $|\phi - \phi'|$  and  $\theta_s$ . Corresponding  $|\phi - \phi'|$  coordinates for different  $\theta_s$  are depicted in Fig. 4(a)-(c). Note that, for given  $\beta$ , the geodesic distance  $R$  in Eq. (11) and the angular distance  $|\phi - \phi'|$  increase as  $\cos \theta_s$  or  $ka$  ( $\underline{a}$  fixed) decreases (see also Figure 2). For example, at  $\beta = 1$ , one has  $R \approx 2.7$  ( $\theta_s = 0^\circ$ ),  $R \approx 3.2$  ( $\theta_s = 30^\circ$ ),  $R \approx 6.8$  ( $\theta_s = 60^\circ$ ),  $R \approx 28$  ( $\theta_s = 80^\circ$ ) for  $ka = 100$  ( $a = 10$  meter), and  $R \approx 4.6$  ( $\theta_s = 0^\circ$ ),  $R \approx 5.6$  ( $\theta_s = 30^\circ$ ),  $R \approx 11.7$  ( $\theta_s = 60^\circ$ ), for  $ka = 20$  ( $a = 10$  meter), where  $R$  is in meters. Since, for fixed  $ka$  and fixed  $\beta$ , there is the simple relation  $|a_1 G(a_1)| = |a_2 G(a_2)|$ , where  $G(a_1)$  and  $G(a_2)$  are calculated from Eqs. (8) or (13) with  $a = a_1$  and  $a = a_2$ , respectively, one need evaluate the field for only one combination of  $k$  and  $\underline{a}$ . Then use of the relation above yields the field for arbitrary radius provided that  $(ka)$  is unchanged. First we shall discuss the accuracy and range of validity of the various formulations in Section C.

In Fig. 4, we have obtained the field magnitudes for various  $ka$  and  $\theta_s$  based on the ray plus canonical integral representation in Eq. (8) which serves as an accurate and convenient reference solution. When  $ka = 20$  and  $\theta_s = 60^\circ$ , then  $k_t a = ka \cos \theta_s = 10$ , so that one may ask whether the requirement of large  $ka \cos \theta_s$  is satisfied here. From experience with the two-dimensional case,<sup>1</sup> confidence is established that the asymptotic formulations in Section C are applicable for these parameter values and hence for the others listed above. The number of rays to be included in Eq. (8) is determined by the inequality Eq. (12), while the canonical integral  $I_N(\beta)$  in Eq. (9b) is evaluated numerically. A tabulation of  $I_N(\beta)$  may be found in Reference 5. As in the two-dimensional case,<sup>1</sup> the choice of  $\Delta$  in Eq. (12) is not very critical; we have used  $\Delta \approx 2$  throughout. Because of the overlap of the curves for various  $N$  (see Fig. 4), switching from one formulation to the other is performed smoothly.

The mix of  $(N+1)$  rays and  $M$  whispering gallery modes in Eq. (13) provides the most intriguing and physically appealing formulation. As noted in Section C-4, the whispering gallery mode representation Eq. (14) is restricted to the lower order W. G. modes since we have used the Fock asymptotic approximations<sup>1</sup> which are valid only for  $\nu \approx k_t a$  and  $k_t a$  large. The eigenvalues  $\nu_m$  in the  $\nu$ -plane and  $t_m$  in the  $t$ -plane are related by  $\nu_m = k_t a + (k_t a/2)^{1/3} t_m$ , where the  $t_m$  ( $m = 1, 2, \dots, M$ ) are the zeros of  $Ai'(t)$ .<sup>7</sup> Note that the  $t_m$  are negative real, and that their magnitude increases with increasing  $m$ , the order of the mode. Therefore, the restriction  $\nu_m \approx k_t a$  required for applying the Fock asymptotic approximations is violated at a certain number  $(M+1)$ .

This limitation did not arise in the two-dimensional case,<sup>1</sup> where we have employed the uniform asymptotic approximations, which are valid for the entire range of

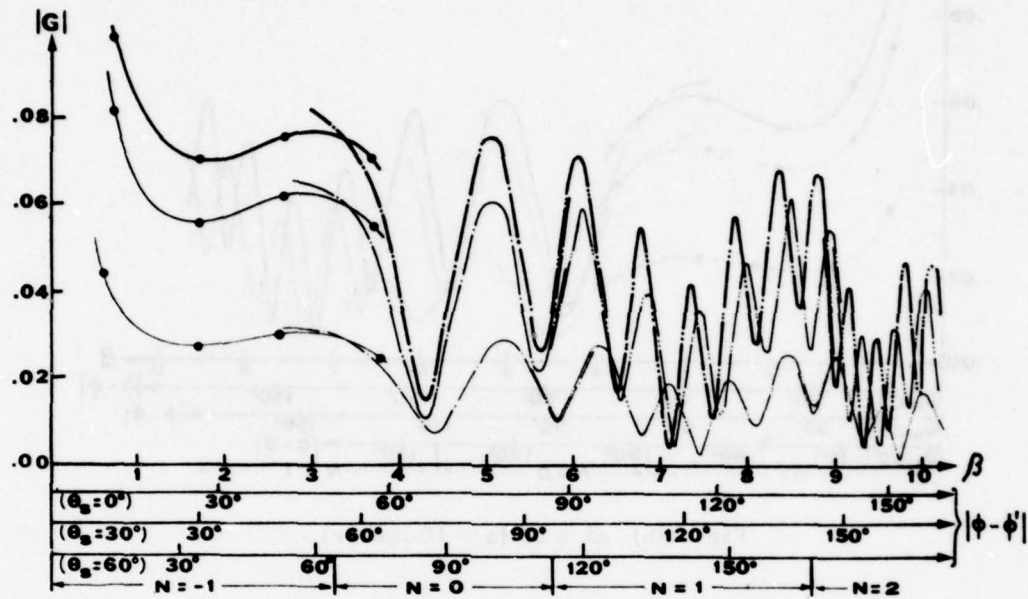
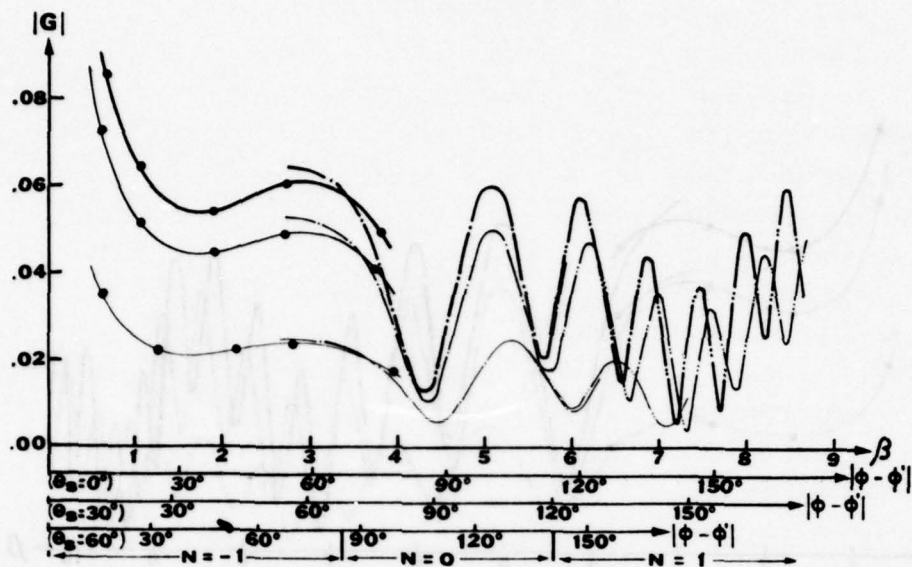
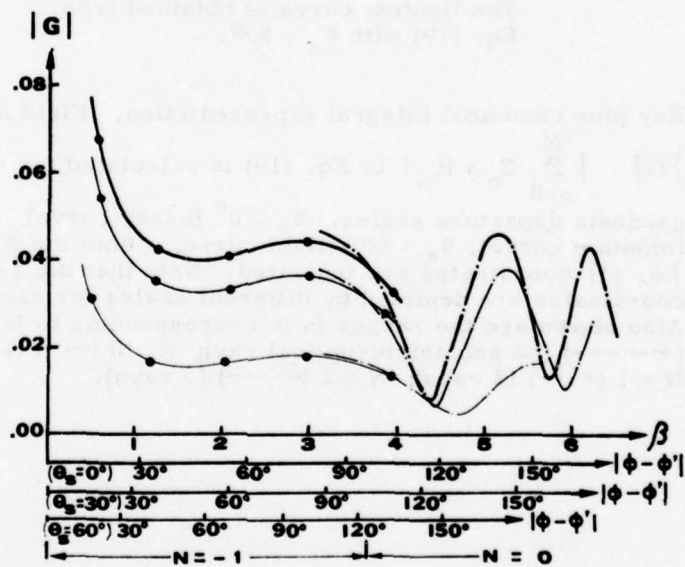


Fig. 4(a)  $ka = 100$  ( $a = 10$  meter,  $f = 477.5$  MHz).  
The lightest curve is obtained from  
Eq. (19) with  $\theta_s = 80^\circ$ .

Fig. 4. Ray plus canonical integral representation. Field magnitude

$|G| = \left| \sum_{n=0}^N G_n + R_N \right|$  in Eq. (19) is calculated for various geodesic departure angles.  $\theta_s = 0^\circ$  (heavy curve),  $\theta_s = 30^\circ$  (medium curve),  $\theta_s = 60^\circ$  (light curve). Both the  $\beta$  and  $|\phi - \phi'|$  coordinates are indicated. Note that the  $|\phi - \phi'|$  coordinates are depicted by different scales for each  $\theta_s$ . Also shown are the ranges in  $\beta$  corresponding to  $N = -1$  (---) (no geometric-optical ray);  $N = 0$  (---) (1 ray);  $N = 1$  (---) (2 rays);  $N = 2$  (---) (3 rays).

Fig. 4(b)  $ka = 50$  ( $a = 10$  meter).Fig. 4(c)  $ka = 20$  ( $a = 10$  meter).

$\nu$  and large  $ka$ , instead of the Fock approximations to obtain asymptotic expressions for the W. G. modes. But to obtain the three W. G. modes from the two-dimensional ones, the additional integration in Eq. (1) is required and can be carried out conveniently (see Eq. (3)) only when the two-dimensional field constituent has the form in Equation (2). Such a form is exhibited by the two-dimensional W. G. modes obtained with the Fock approximations but not by those obtained with the uniform approximations ( $D$  in Eq. (2) is in general a complicated function of  $k$ ). Thus, while it was possible (for  $\gamma \geq 3.1$ ) to represent the two-dimensional field by a single ray plus an appropriate number of W. G. modes,<sup>1</sup> this cannot in general be done with our results for the three-dimensional field. Note, however, that when the observation point is on the transverse cross-section (i.e.,  $\theta_s = 0^\circ$ ), Eq. (2) still holds for the uniform expression of the W. G. modes, thereby making it possible to represent the field by a single ray plus W. G. modes for this special case.

The fields computed from the ray plus whispering gallery mode representation in Eq. (13) and from the near field form in Eq. (15) are compared with the reference solutions (solid curves) obtained from Eq. (8) in Figure 5(a)-(c). The ray plus whispering gallery mode representation is invalid in the near field, when the left-hand side of the inequality in Eq. (12) is negative; here, one may use the near field form. From each of the curves in Figs. 5(a)-(c), it is seen that changeover from the near field form

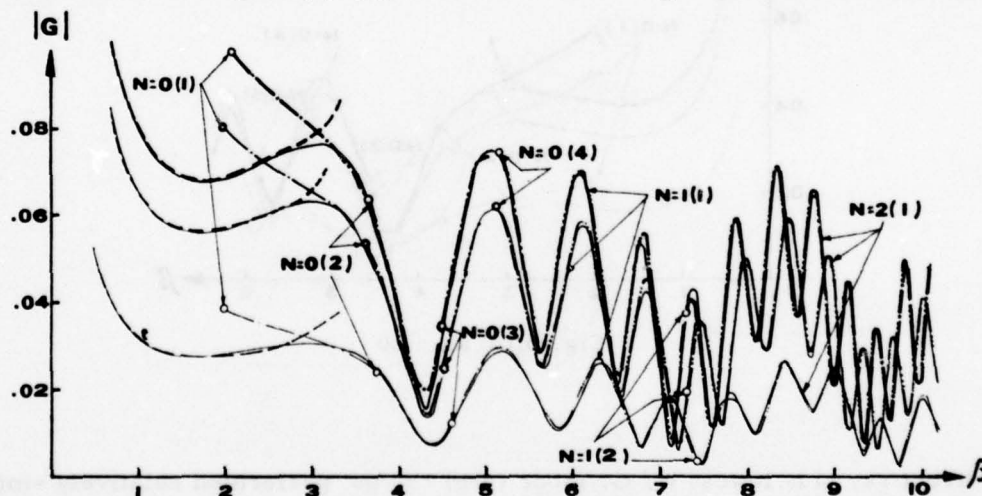
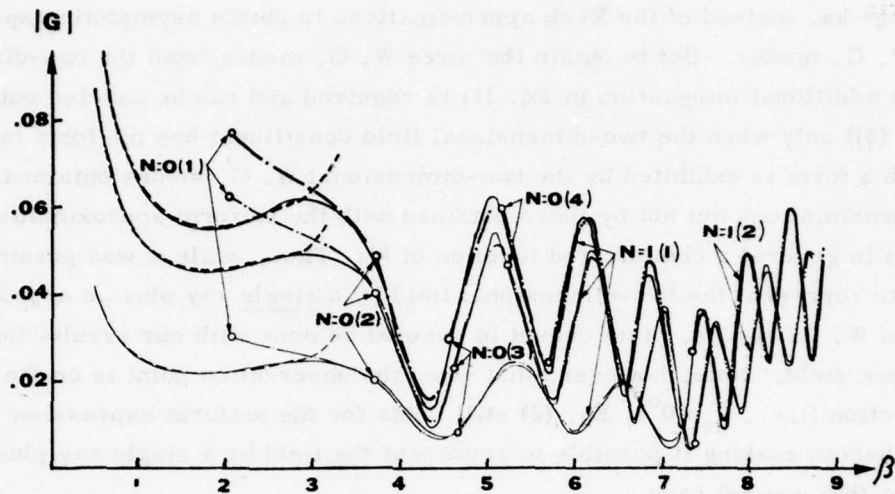
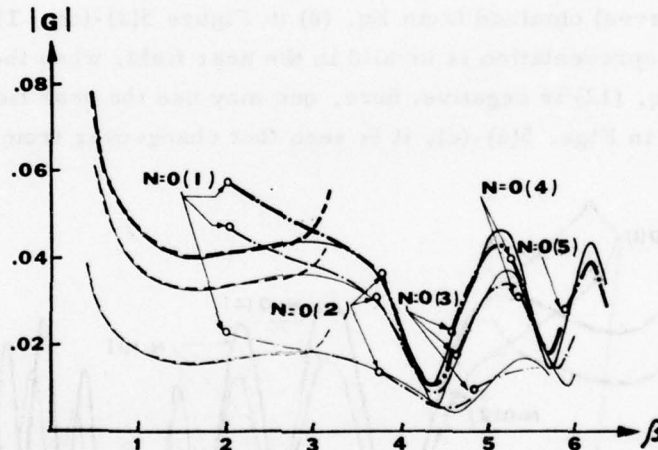


Fig. 5(a)  $ka = 100$  ( $a = 10$  meter). The lightest curve is calculated from Eq. (27) with  $\theta_s = 80^\circ$ .

Fig. 5. Ray plus whispering gallery mode representation in Eq. (27) and near field form in Eq. (32) (---). The reference solution (solid curve) is calculated from Eq. (19). The numbers indicated along the curves should be read as follows: for example,  $N=1(2)$  denotes the range wherein the direct and singly reflected ray plus 2 whispering gallery modes are applicable, while the circles indicate the starting points of the relevant intervals. The heavy, medium and light curves are obtained from Eq. (27) with  $\theta_s = 0^\circ$ ,  $\theta_s = 30^\circ$  and  $\theta_s = 60^\circ$ , respectively.

Fig. 5(b)  $ka = 50$ Fig. 5(c)  $ka = 20$ 

to the direct ray plus lowest W. G. mode form can be performed relatively smoothly at the crossover. In the near field computation, 11 terms in the series expansion have been used since this has yielded the smoothest changeover. It is also seen to be possible to represent the field accurately by only the fundamental W. G. mode plus a proper number of rays determined by the inequality. These excellent approximations are valid if the projection of the observation point on the transverse cross-section is

characterized by parameter values away from the region of coalescence of a modal pole and a saddle point in the remainder integral  $R_{MN}$  (see Ref. 1) which corresponds to the geometrical contact of ray and modal caustic as shown in Figure 3. The jumps in each of the curves in Figs. 5(a)-(c) correspond to these points and can be avoided by resorting to a Fresnel integral formulation of the asymptotic field in this transition region. However, as in the two-dimensional case,<sup>1</sup> we have found that continuity can be established by using the simple result for coincident pole and saddle point and then resorting to a perturbation expansion.

Next, we discuss the field behavior as a function of various physical parameters. To explore the dependence on frequency and on the geodesic departure angle, it is more convenient to plot  $|G|$  vs.  $R$  curves, where  $R$  is the geodesic distance between source and observation point defined in Equation (11). Figure 6 exhibits the field for

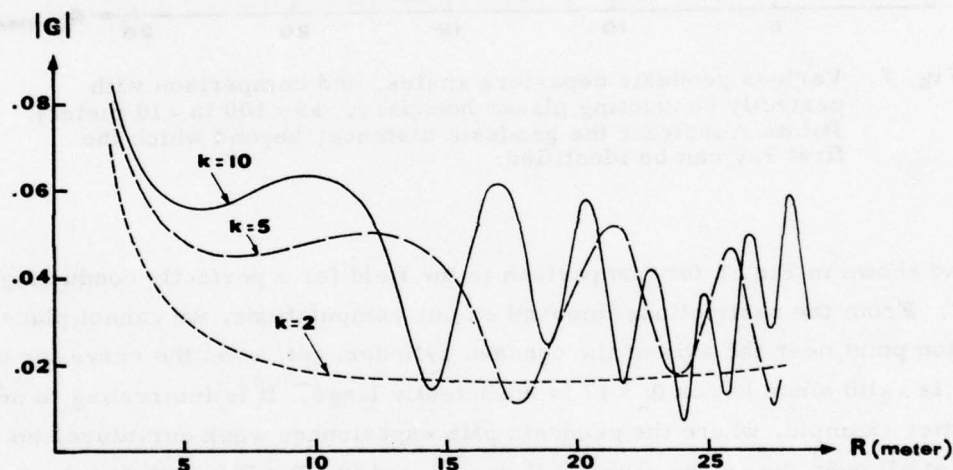


Fig. 6. Frequency dependence of the field.  $a = 10$  meter,  $\theta_s = 30^\circ$ .

$k = 2\pi/\lambda = 2, 5, 10$ , with  $a$  and  $\theta_s$  fixed at 10 meter and  $30^\circ$ , respectively, while in Fig. 7,  $\theta_s$  takes on values of  $0^\circ, 30^\circ, 60^\circ, 80^\circ$ , and  $ka = 100$  ( $a = 10$  meter) is fixed. It is evident from the figures that the near field decreases with the frequency and with increasing  $\theta_s$ . This behavior indicates that, for an isotropic radiator, the near field ( $R \lesssim 10$ ) is enhanced on the transverse cross-section, where  $\theta_s = 0^\circ$ , and at higher frequencies. Interference effects between rays are weakened for decreasing  $k$  and for increasing  $\theta_s$ , as seen from the decreasing number of oscillations in the curves. The accompanying shifts of the field maxima and minima to the right indicate that when observed along the geodesic path, the ray field develops more quickly at high frequencies and at small geodesic departure angles than at lower frequencies and at larger geodesic angles.

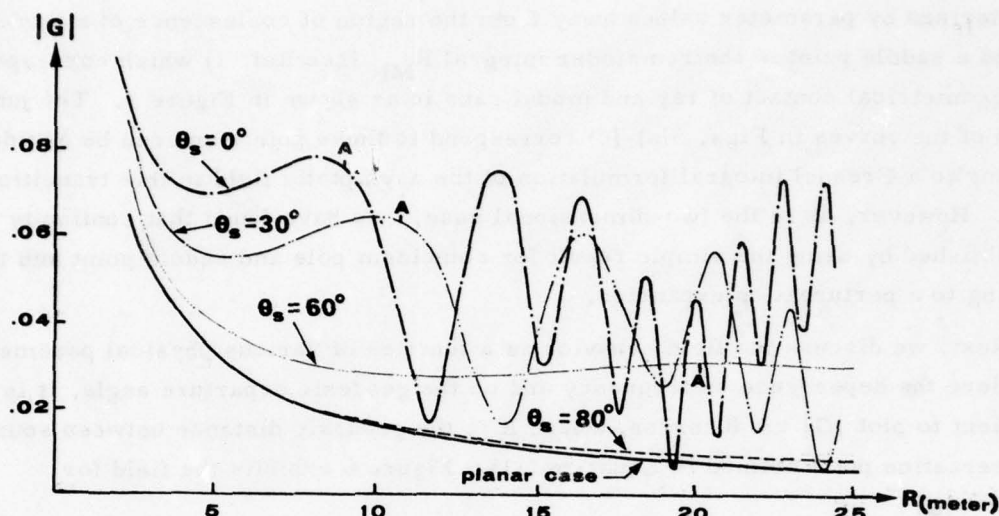


Fig. 7. Various geodesic departure angles, and comparison with perfectly conducting planar boundary.  $ka = 100$  ( $a = 10$  meter). Points A indicate the geodesic distance, beyond which the first ray can be identified.

Also shown in Fig. 7 for comparison is the field for a perfectly conducting planar boundary. From the restrictions imposed on our computations, we cannot place the observation point near the axis of the concave cylinder; yet, even the curve for  $\theta_s = 80^\circ$  ( $ka = 100$ ) is valid since  $ka \cos \theta_s \approx 17$  is sufficiently large. It is interesting to note that in this latter example, where the geodesic path experiences weak curvature, no ray develops at all over the entire range of  $R$  considered ( $0 \leq R \leq 56$ ), and that the field resembles the one on a perfectly conducting planar boundary over a substantial distance.

U. S. Army Research Office  
DAHC 04-75-G-0152

L. B. Felsen and T. Ishihara

#### REFERENCES

1. T. Ishihara, L. B. Felsen and A. Green, "High Frequency Fields Excited by a Line Source Located on a Perfectly Conducting Concave Cylindrical Surface," to be published in IEEE Transactions on Antennas and Propagation.
2. T. Ishihara and L. B. Felsen, "High Frequency Fields Excited by a Line Source Located on a Concave Cylindrical Impedance Surface," to be published in IEEE Transactions on Antennas and Propagation.
3. L. B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," Prentice Hall, Inc., Englewood Cliffs, New Jersey, p. 382 (1973).
4. L. B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," Prentice Hall, Inc., Englewood Cliffs, New Jersey, p. 382 (1973).

5. T. Isihara, "Tabulation of a Canonical Integral Occurring in the Theory of Wave Propagation along a Concave Cylindrical Impedance Surface," submitted for publication.
6. G. Hasserjian and A. Ishimaru, "Currents Induced on the Surface of a Conducting Circular Cylinder by a Slot," J. Res. Nat. Bur. Stand., Vol. 66D, No. 3, pp. 335-365 (May-June 1962).
7. W. Wasylkiwskyj, "Exact and Quasi-Optic Diffraction within a Concave Cylinder," IEEE Transactions on Antennas and Propagation, Vol. AP-23, No. 4, pp. 480-492 (July 1975).
8. M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions," Dover Publications, Inc., New York, p. 478 (1972).

## PERIODIC STRUCTURE GTD FOR ANALYSIS OF MUTUAL COUPLING IN ARRAYS ON CONCAVE SURFACES\*

H. Ahn and A. Hessel

A. Introduction

Arrays on concave surfaces form an integral part of electronically scanned space-fed (feed-through) lenses of the dome antenna,<sup>1</sup> or the multibeam bootlace type.

This report surveys the development of the so-called Periodic Structure Geometric Theory of Diffraction (PSGTD) for evaluation of coupling (scattering) coefficients in "collector" arrays of aperture elements concave surfaces. The method of analysis and some partial results for far mutuals were briefly described in JSTAC Report No. 42. We have since carried out the development of the relevant transition function for the evaluation of the near zone coupling. For the sake of clarity and continuity we shall review the entire development.

The development of PSGTD has been motivated by the low efficiency (or the poor convergence) in concave geometry of the usual method of evaluating the mutual coupling, which relies on numerical calculation of the mutual admittance coefficients, followed by inversion of the admittance matrix. The convergence difficulties arise from the very slow decay of mutual interaction in concave geometry, as compared to that in arrays on planar or convex surfaces.<sup>2</sup> The reason for such a slow decay is the presence of line of sight between each of the array elements in concave geometry. To circumvent this difficulty and to develop a suitable alternative approach, it is necessary to depart from a canonical problem associated with a conducting surface and begin the analysis with an appropriate prototype problem, specifically, that of coupling coefficients in an angularly periodic, circularly cylindrical, concave array (Fig. 1). In that case, the high frequency asymptotic evaluation of fields or coupling coefficients gives rise to a variant of the geometric theory of diffraction, referred to as Periodic Structure GTD, because the relevant canonical launching, reception, and reflection problems are associated with periodic planar arrays, rather than with conducting planar surfaces, as in the standard GTD practice.

The virtues of the PSGTD in application to arrays on concave surfaces are:

- a) The results are generalizable to slow variation of curvature and periodicity.
- b) The method employs planar phased array results as canonical problems for launching, attachment and reflection coefficients.
- c) It permits exploitation of planar phased array computer programs, and does not require inversion of large matrices.
- d) It furnishes simple physical explanation of the characteristic features of numerical results.

---

\*This material was presented at the 1978 URSI General Assembly in Helsinki, Finland.

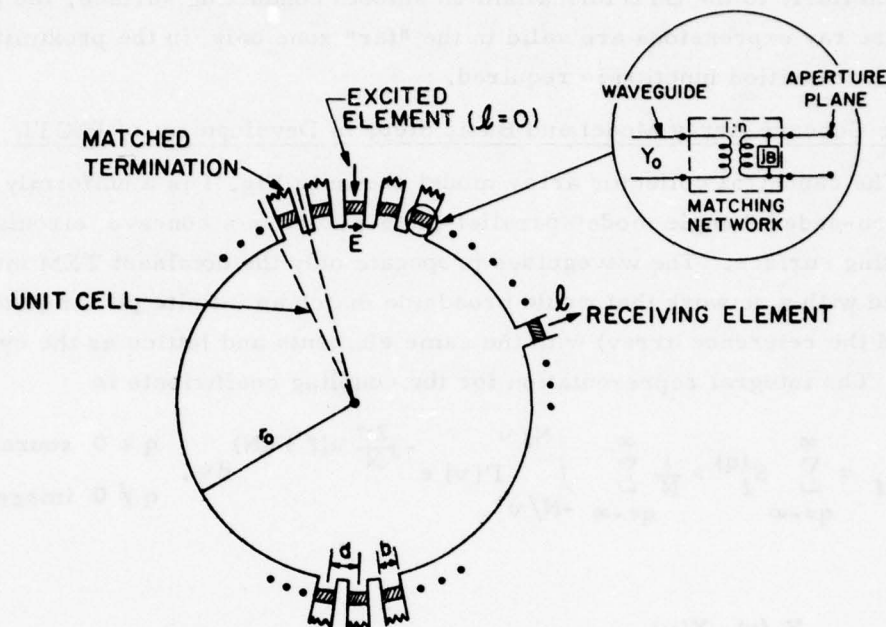


Fig. 1. Cylindrical collector array.

As will be seen subsequently, the periodic structure rays travel along straight lines and their geometric optical trajectories are identical with those occurring when both the source and the observation point are located on the unperforated, concave, conducting array support surface. These ray fields represent local plane waves, radiated by the excited element and observed in the receiving element, in a matched terminated array environment. Their launching and reception (attachment) coefficients are no longer those on a planar conducting surface, but correspond to planar periodic arrays appropriate to the respective environments of the transmitting or the receiving element on the concave surface. In passing from the former to the latter, the ray fields are weighted by the angle dependent launching and attachment coefficients which are proportional to the transmitting and receiving element patterns in infinite planar arrays with elements and lattice corresponding to the respective local array geometries on the concave surface. As in the case of a smooth concave conducting surface, there appears here also a direct ray and a series of multiple reflected ray contributions. The reflection coefficients of the latter are no longer from a perfectly conducting surface, but correspond to a plane wave incidence, in the ray direction, on a periodic planar array with a lattice appropriate to that of the local environment surrounding the point of reflection.

Similarly to the GTD formalism on smooth conducting surface, the periodic structure ray expressions are valid in the "far" zone only; in the proximity of the source a transition function is required.

### B. The Concave Array Model and Basic Steps in Development of PSGTD

The canonical collector array model shown in Fig. 1 is a uniformly spaced array of  $N$  open-ended "single mode" parallel-plate-guides in a concave, circularly cylindrical conducting surface. The waveguides propagate only the dominant TEM mode and are equipped with a network that would broadside match an infinite planar phased array (termed the reference array) with the same elements and lattice as the cylindrical array. The integral representation for the coupling coefficients is

$$S_\ell = \sum_{q=-\infty}^{\infty} S_\ell^{(q)} = \frac{1}{N} \sum_{q=-\infty}^{\infty} \int_{-N/\nu}^{N/\nu} \Gamma(\nu) e^{-j\frac{2\pi}{N} \nu(\ell + qN)} d\nu, \quad \begin{array}{l} q = 0 \text{ source term} \\ q \neq 0 \text{ image terms.} \end{array} \quad (1)$$

Here

$$\Gamma(\nu) = \frac{Y_p(o) - Y(\nu)}{Y_p^*(o) + Y(\nu)} \quad (2)$$

and

$$Y(\nu) = j\sqrt{\frac{\epsilon_o}{\mu_o}} \frac{b}{d} \sum_{m=-\infty}^{\infty} \left( \frac{\sin \nu_m b/2r_o}{\nu_m b/2r_o} \right)^2 \frac{J_{|\nu_m|}(kr_o)}{J'_{|\nu_m|}(kr_o)} ; \quad \nu_m = \nu + mN. \quad (3)$$

In Eqs. (1) and (2)  $\Gamma(\nu)$  is the active reflection coefficient corresponding to a Floquet index  $\nu$  in the wedge-shaped unit cell of Fig. 1, and  $Y(\nu)$  is the corresponding active aperture admittance.  $Y_p(o) = G_p(o) + jB_p(o)$  is the active broadside-scan admittance of the unmatched reference array.

Steps in the asymptotic evaluation of  $S_\ell^q$ :

1. Debye approximation for  $J_{|\nu_m|}(kr_o) J'_{|\nu_m|}(kr_o)$ .
2. Expansion of  $\Gamma(\nu)$  into Neumann series.
3. Stationary phase evaluation term-by-term of the series yields periodic structure ray contributions to "far" zone mutuals.
4. For near neighbors: Airy function approximation for  $J_{\nu_o}$  and contour deformation yield the transition function.

### C. Periodic Structure Ray Fields

The stationary phase (or the periodic structure ray) expressions for the coupling coefficients  $S_l$  are shown in Fig. (2) along with a number of typical rays. One notes that  $S_l$  is a superposition of various ray contributions  $S_l^{qn}$ . Numerical results show that only a small number of rays is required for a sufficient accuracy. On the basis of the phase term one concludes that the ray paths are straight lines, identical to those for a magnetic line source located in the center of the aperture of the excited element, for observation point in the center of the aperture of the receiving element. Each ray is described by three indices. The subscript  $l$  denotes the receiving element, the superscript  $q \neq 0$ , indicates the image source from which the ray originates (such rays sweep through an angle larger than  $180^\circ$ , subtended at the center before reaching the

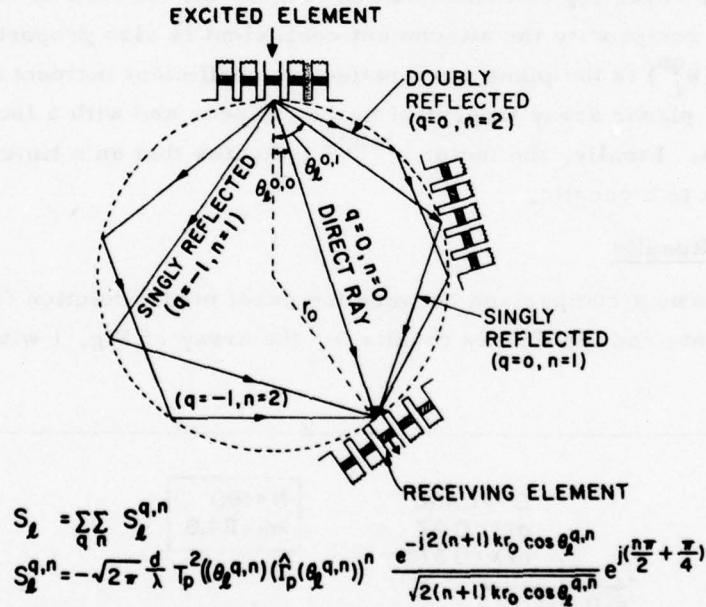


Fig. 2. Periodic structure ray.

receiving element (positive (or negative) values of  $q$  correspond to rays traveling clockwise (or counterclockwise)). The index  $n$  denotes the number of reflections of the ray during its passage from the transmitting to the receiving element. Figure 2 shows the direct, singly and doubly reflected  $q=0$  rays and the singly and doubly reflected  $q=-1$  rays. One notices that, in view of the circular cylindrical geometry, the ray exit angle  $\theta_{l,q,n}$  at the transmitting element equals to the incidence angle of this ray, both with respect to the appropriate local normals. When the receiving element approaches the

source  $\theta_l^{qn}$  approaches  $90^\circ$  and the ray path becomes too short for the ray concept to be valid. In formal terms, the Debye approximation of  $J_\nu$  breaks down at the stationary point. The appearance of the factor  $T_p^2$  and  $\hat{\Gamma}_p^2$  is the basic feature of PSGTD rays and is understood as follows. The factor  $T_p(\sin \theta_l^{qn})$  is the transmission coefficient of an infinite planar phased array (the reference array) located tangentially at the excited element site (Fig. 2) and scanned to an angle  $\theta_l^{qn}$  off broadside. The elements and lattice of this array are appropriate to the local geometry of the cylindrical array in the vicinity of the excited element. From phased array theory  $T_p(\sin \theta_1)$  is also the pattern in the ray direction of the transmitting element in the above match-terminated array environment. The appearance of  $T_p^2(\sin \theta_l^{qn})$  corresponds to the product of the launching and attachment coefficient. The latter is recovered from the reception problem of a plane wave incident in the ray direction on a planar array tangential to the cylinder at the receiving element location (Figure 2). In view of circular cylindrical geometry and reciprocity the attachment coefficient is also proportional to  $T_p(\sin \theta_l^{qn})$ . The factor  $\hat{\Gamma}_p(\sin \theta_l^{qn})$  is the plane wave reflection coefficient incident in the ray direction on an infinite planar array tangential to the cylinder and with a local lattice at the point of reflection. Finally, the factor  $e^{jn\pi/2}$  indicates that an  $n$  times reflected ray is  $n$  times tangent to a caustic.

### Numerical Results

Figure 3 shows a comparison between the exact modal solution for magnitude of coupling coefficients and the PS ray results for the array of Fig. 1 with  $kr_0 = 84.6$

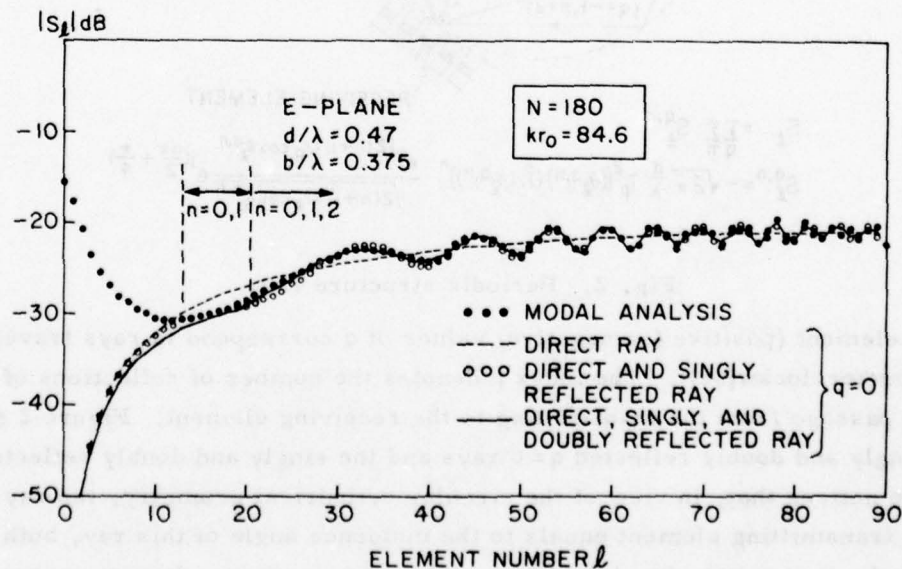


Fig. 3. Comparison of modal and ray analysis.

and with the total number of 180 elements spaced  $0.47\lambda$  apart. The exact solution indicates a  $|S_\ell|$  curve falloff which follows initially that in a planar reference array. Gradually, however, the decay rate of  $|S_\ell|$  versus the element number decreases and the curve passes through a dip with a subsequent rise, accompanied later by a distinct ripple.

All ray contributions belong to the  $q=0$  family. The direct ray (Fig. 3) contribution (dashed) sets the average level of the curve past the depression and departs from the exact solution near the source for  $\ell \leq 14$ . Superposition of the direct and the singly reflected ray yields the rimmed circles and demonstrates that the characteristic ripple is caused predominantly by the interference between the direct and the singly reflected ray fields.

The rise of the curve past the dip region can be simply explained in terms of the direct ray result. By the phased array theory and geometry

$$|S_\ell| \sim \frac{|T_p(\sin \theta_\ell^{oo})|}{\sqrt{kr}} = \frac{\cos \theta_\ell^{oo} (1 - |\Gamma_p(\sin \theta_\ell^{oo})|^2)}{\sqrt{2kr_0} \sqrt{\cos \theta_\ell^{oo}}} = \frac{\sqrt{\cos \theta_\ell^{oo}} (1 - |\Gamma_p(\sin \theta_\ell^{oo})|^2)}{\sqrt{2kr_0}}$$

as  $\ell$  increases,  $\theta_\ell^{oo}$  decreases and  $\sqrt{\cos \theta_\ell^{oo}}$  increases. Also, as  $\theta_\ell^{oo}$  decreases the planar active array reflection coefficient likewise decreases, since the elements are broadside matched. As a result  $|S_\ell|$  increases as the receiving element approaches the diametrically opposite position. Figure 4 shows the improvement in the ray result

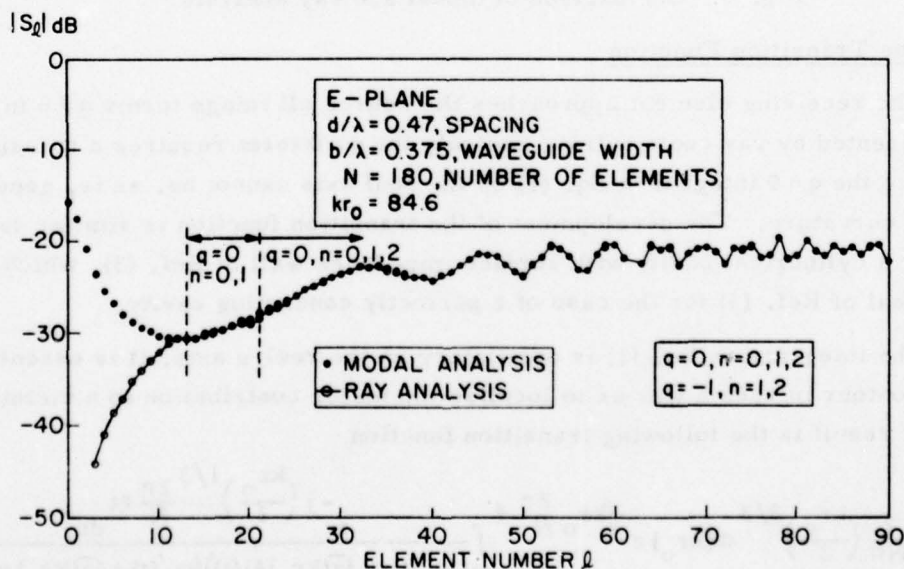


Fig. 4. Comparison of modal and ray analysis.

when five most significant rays,  $q = 0, n = 0, 1, 2$  and  $q = -1, n = 1, 2$ , are incorporated. Figure 5 indicates a similar agreement for an array with about half the radius ( $kr_0 \sim 42$ ). Again, the correlation is excellent beyond the center of the dip. The phase of  $S_\ell$  calculated by the ray approach tracks that of modal analysis to within a few degrees.

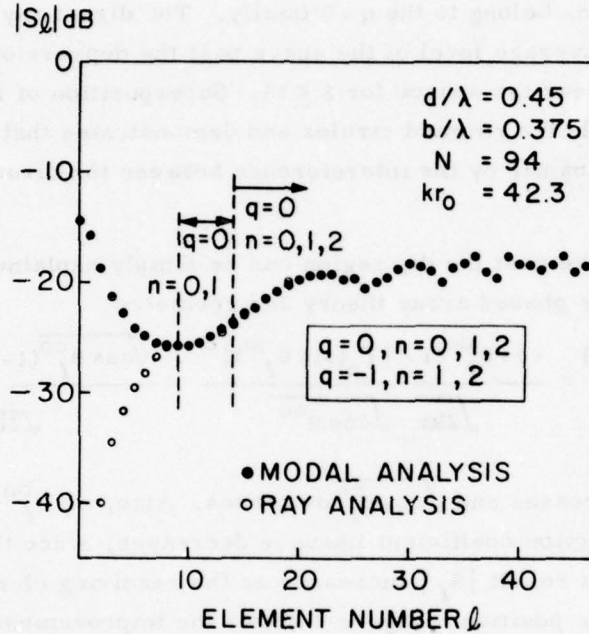


Fig. 5. Comparison of modal and ray analysis.

#### D. Near Zone Transition Function

When the receiving element approaches the source all image terms  $q \neq 0$  in Eq. (1) will be represented by ray contributions and only the  $q = 0$  term requires a transition function, since the  $q = 0$  integral in Eq. (1) on the real axis cannot be, as is, generalized to a variable curvature. The development of the transition function is similar to that employed for a cylindrical cavity with surface impedance wall in Ref. (3), which in turn follows an ideal of Ref. (4) for the case of a perfectly conducting cavity.

Since the integrand in Eq. (1) is oscillatory on the real  $v$  axis, it is essential to deform the contour in such a way as to localize the major contribution to a vicinity of  $v = kr_0$ . The result is the following transition function

$$S_\ell \sim -\frac{2}{N\pi} \left(\frac{kr_0}{2}\right)^{2/3} \alpha(kr_0) e^{-jkr_0 \frac{2\pi}{N} \ell} \int \frac{e^{-j\left(\frac{kr_0}{2}\right)^{1/3} \frac{2\pi}{N} \ell t}}{P(Ai'(t) + j\bar{z}(kr_0)Ai(t))(w_1'(t) + j\bar{z}(kr_0)w_1(t))} dt \quad (4)$$

$$\alpha(kr_o) = 2 \left( \frac{\sin kb/2}{kb/2} \cdot \frac{G_p(0)}{Y_p^*(0) + Y_p'(kr_o)} \right) \quad (5)$$

$$\bar{z}(kr_o) = -\sqrt{\frac{\epsilon_o}{\mu_o}} \frac{b}{d} \left( \frac{\sin kb/2}{kb/2} \right)^2 \frac{1}{Y_p^*(0) + Y_p'(kr_o)} \cdot \left( \frac{kr_o}{2} \right) \quad (6)$$

$$Y_p'(kr_o) = j\sqrt{\frac{\epsilon_o}{\mu_o}} \frac{b}{d} \sum_{m \neq 0} \left[ \frac{\sin \left[ \frac{kb}{2} \left( 1 + \frac{m\lambda}{d} \right) \right]}{\left( 1 + \frac{m\lambda}{d} \right) \frac{kb}{2}} \right]^2 \frac{1}{\sqrt{\left( 1 + \frac{m\lambda}{d} \right)^2 - 1}} \quad (7)$$

where the integration variable  $t$  arises from the usual transformation

$$v = kr_o + \left( \frac{kr_o}{2} \right)^{1/3} t \quad (8)$$

The details of the contour are given in Figure 6. Figures 7(a) and 7(b) show a comparison between the exact modal result (black circles), the transition function for cylinder

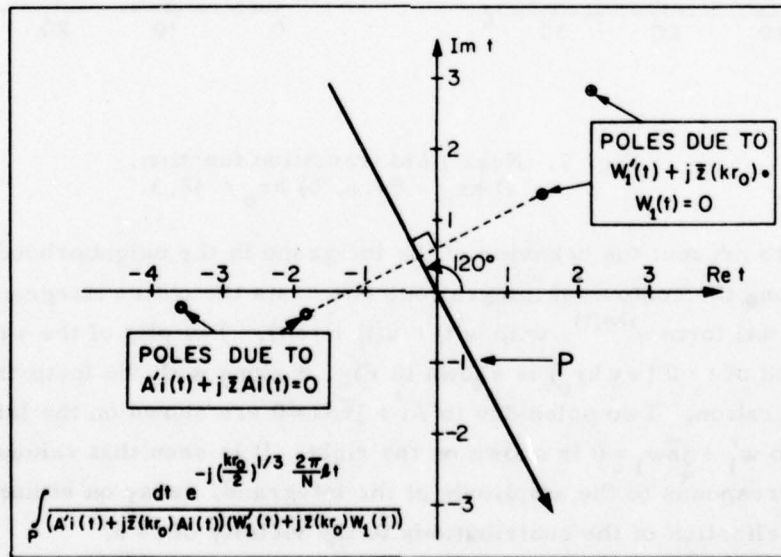


Fig. 6. Contour for near zone transition function.

radii  $kr_0 = 84.6$  and  $kr_0 = 42.3$  and the planar result. It is seen that the agreement is quite good, but for the first neighbor ( $l = 1$ ) one may wish to use the planar result for better accuracy. On the other hand, there is a fair amount of overlap in the dip region between the ray results of Figs. 4 or 5 and the transition function of Figures 7(a) and 7(b).

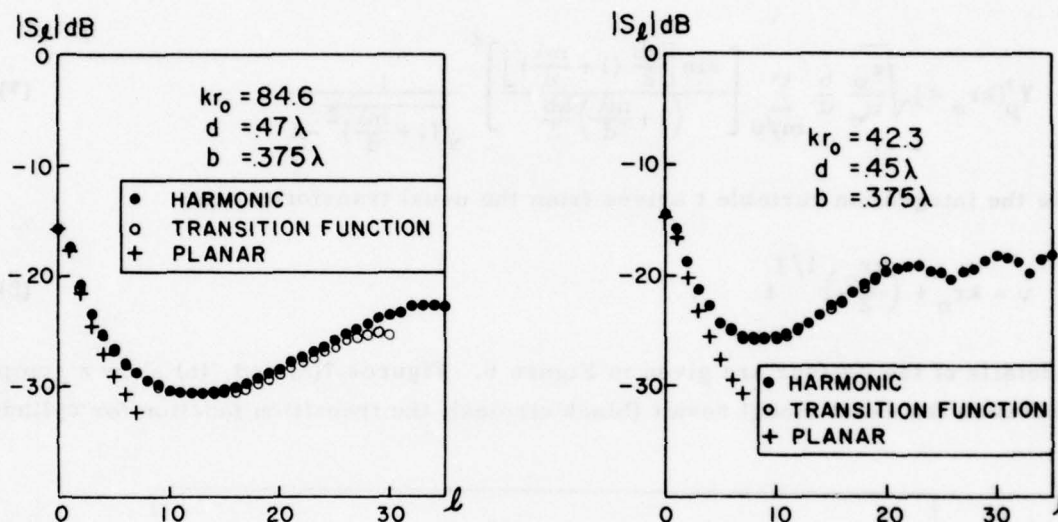
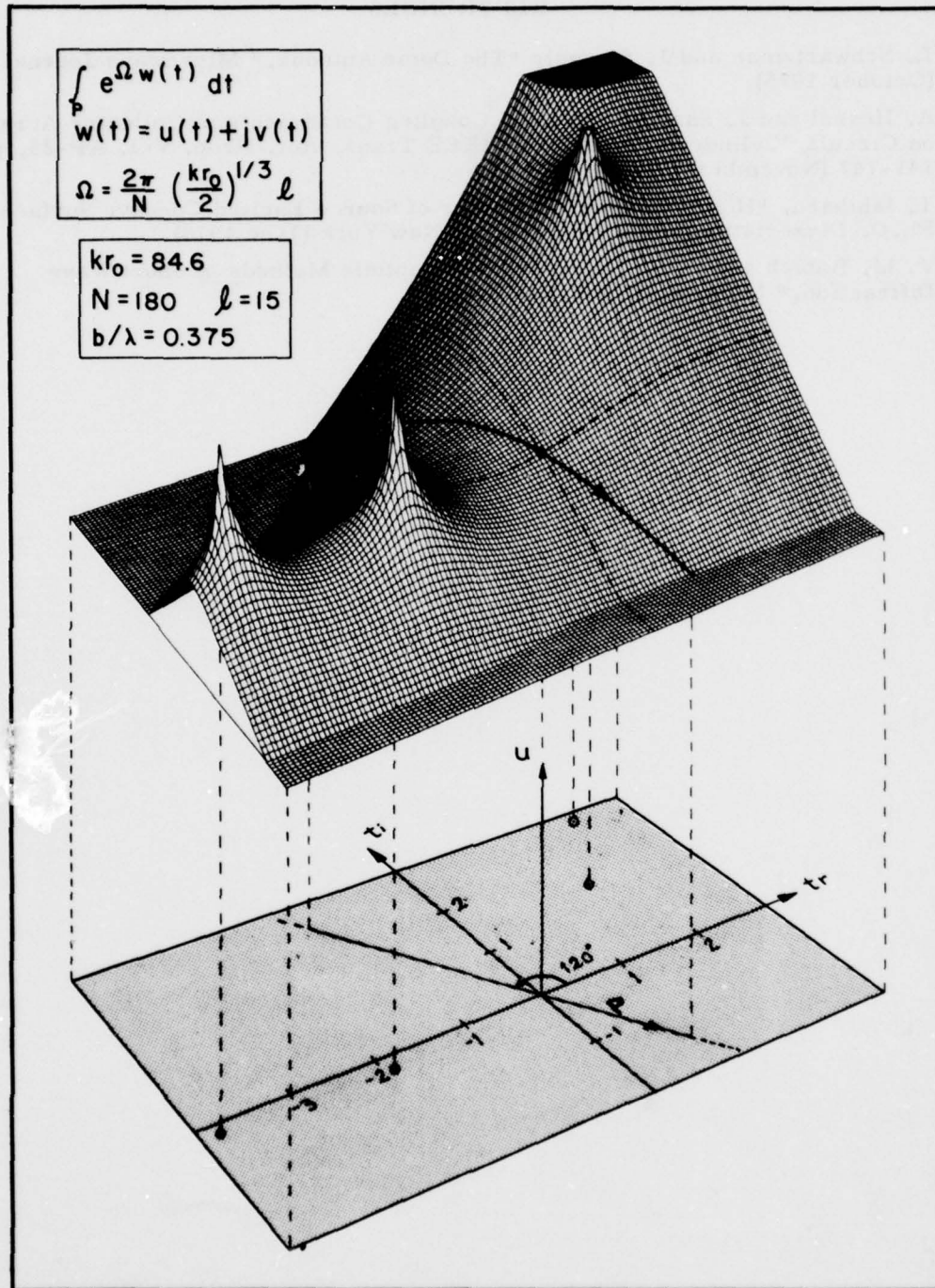


Fig. 7. Near field transition function.  
a)  $kr_0 = 84.6$ , b)  $kr_0 = 42.3$ .

In order to present the behavior of the integrand in the neighborhood of  $t = 0$  and particularly along the contour of integration, one casts the entire integrand in Eq. (4) into an exponential form  $e^{\Omega w(t)}$ , with  $w(t) = u(t) + jv(t)$ . The plot of the surface  $u(t)$  in the neighborhood of  $t = 0$  ( $v = kr_0$ ) is shown in Fig. 8 along with the footprint of the contour of integration. Two poles due to  $Ai' + j\bar{z}Ai = 0$  are shown on the left and the first pole due to  $w_1' + j\bar{z}w_1 = 0$  is shown on the right. It is seen that values of  $u(t)$ , where  $e^{u(t)}$  corresponds to the amplitude of the integrand, decay on either side of  $t = 0$ , indicating a localization of the contributions to the vicinity of  $t = 0$ .

Fig. 8. Surface  $u(t)$  near  $t=0$ .

## REFERENCES

1. L. Schwartzman and J. Stangel, "The Dome Antenna," *Microwave Journal* (October 1975).
2. A. Hessel and J. Shapira, "Mutual Coupling Coefficients in Collector Arrays on Circular Cylindrical Surfaces," *IEEE Trans. Ant. Prop.* Vol. AP-25, pp. 741-747 (November 1977).
3. T. Ishihara, "High Frequency Behavior of Source Excited Concave Surfaces," Ph.D. Dissertation, Polytech. Inst. of New York (June 1978).
4. V. M. Babich and V. S. Buldyrev, "Asymptotic Methods of Short Wave Diffraction," Moscow, U.S.S.R.

## LIMITATIONS ON GAIN VERSUS SCAN FOR A DOME ANTENNA

H. Steyskal, A. Hessel and J. Shmoys

This study considers limitations on the gain and its scan variation in a 2-dimensional dome antenna,<sup>1</sup> as shown in Figure 1. An upper bound on the average

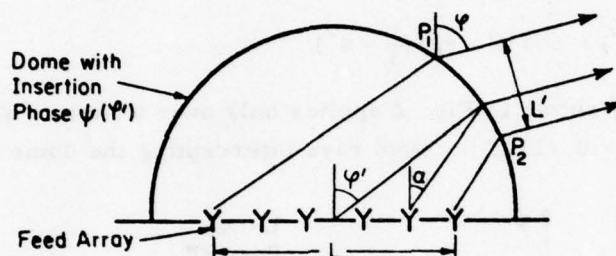


Fig. 1. Two-dimensional dome antenna.  $L$  = feed array length,  $a$  = dome radius,  $\varphi$  = scan direction,  $\alpha$  = local scan angle of feed array,  $\varphi'$  = angular dome coordinate.

gain over the scan sector is derived, which is lower than that given in Ref. 2 and an upper bound on the gain in any particular scan direction is found. Illustrative examples of gain-vs-scan variation are given.

The analysis employs the following assumptions about the antenna system: 1) circular cylindrical, thin dome with a fixed insertion phase distribution  $\psi$  and with a feed array located on the diameter; 2) scan coverage sector  $-90^\circ \leq \varphi \leq 90^\circ$ ; 3) symmetry about zenith; 4) use of entire feed array for all scan directions; 5) specified scan angle limits for feed array, dome collector and dome radiator sides (in this report  $60^\circ$  was used for all three); 6) validity of ray optics; 7) absence of crossed rays; 8) all of the dome surface within the scan range of the feed array is employed.

The feed array phasing is assumed such that the far fields radiated by all elements always add coherently in the desired scan direction. It then can be shown that the following upper bound holds for the average gain (normalized to feed array broad-side gain  $2\pi L/\lambda$ )

$$\bar{g} = \frac{1}{\pi(2\pi L/\lambda)} \int_{-\pi/2}^{\pi/2} G(\varphi) d\varphi \leq \frac{2}{\pi} \eta \sin \alpha_{\max} \quad (1)$$

where  $G$  denotes the dome antenna gain,  $\eta$  the feed array aperture efficiency and  $\alpha_{\max}$  is the maximum permissible feed array scan angle. Thus, scanning the array to  $\alpha_{\max} < 90^\circ$  results in a loss of average gain ( $\approx 13\%$  for  $\alpha_{\max} = 60^\circ$ ).

The scan angle constraint also leads to a lower bound on the normalized dome phase gradient  $\phi(\varphi') = (\lambda/2\pi a) d\psi/d\varphi'$ , which in turn results in an upper bound on  $G(\varphi)$ . The refraction through the dome is given by the ray equation

$$\sin(\varphi - \varphi') = \sin(\alpha - \varphi') + \phi(\varphi') \quad (2)$$

When  $\alpha_{\max} = 60^\circ$ , this leads, for the worst case ( $\varphi = 90^\circ$ ), to

$$\phi(\varphi') \geq \phi_{\min}(\varphi') = \cos \varphi' - \sin(\frac{\pi}{3} - \varphi') \quad (3)$$

The function  $\phi_{\min}(\varphi')$  shown in Fig. 2 applies only over a limited sector of the dome ( $40^\circ \leq \varphi' \leq 80^\circ$  with  $a = 0.71L$ ), because rays intercepting the dome outside this sector

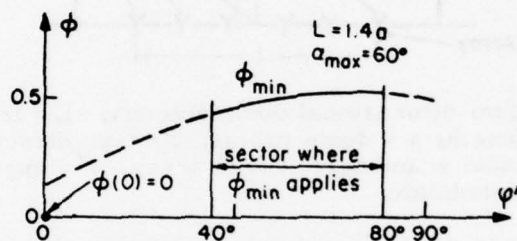


Fig. 2. Minimum phase gradient required not to exceed  $60^\circ$  scan angle on the feed array.

by geometry alone (i.e., independently of phase gradient) correspond to scan angles  $< 60^\circ$ . Another restriction on the phase gradient arises from the required symmetry about zenith which gives  $\phi(0) = 0$ .

For the gain  $G(\varphi)$  to be maximum in a particular scan direction  $\varphi$ , requires (see Fig. 1) that the equivalent aperture  $L'(\varphi)$  be maximized, i.e.,  $\phi$  shall be such that the two intercept points  $P_1$  and  $P_2$  be maximally separated. A systematic use of the conditions on  $\phi$ , together with the requirements that the scan angle off the dome is at most  $60^\circ$ , and that no crossed rays occur, leads to an upper limit function  $L'_{\max}(\varphi)$  and a normalized maximum gain  $g_{\max}(\varphi) = L'_{\max}(\varphi)/L$ . A minimum gain value  $g_{\min}$  can be estimated at  $\varphi = 90^\circ$  by observing that  $L'(90^\circ)$  has a lower bound. Figure 3 shows the function  $g_{\max}(\varphi)$ ,  $g_{\min}(90^\circ)$ , the average  $\bar{g}(\eta = 1, \alpha_{\max} = 60^\circ)$  and the normalized upper limit  $g_p$  given by the projected dome area, for a dome radius  $a = 0.71L$ .

It is noted in Fig. 3, with  $\alpha_{\max} = 60^\circ$ , that there is relatively little difference between  $g_{\max}$  and  $\bar{g}$  in the sector  $40^\circ \leq \varphi \leq 90^\circ$  and between  $g_{\max}$  and  $g_{\min}$  at  $90^\circ$ . This restricts the amount of possible variation in  $G(\varphi)$ . A larger dynamic range is obtained by increasing the relative dome size or relaxing the scan angle limits, as indicated by the curves in Figure 4.

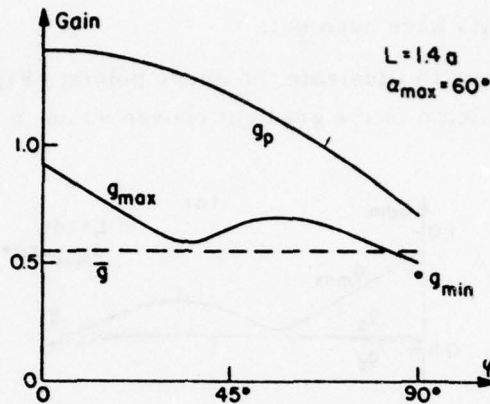


Fig. 3. Gain limits for a dome antenna.  $\bar{g}$  denotes average gain,  $g_{\max}(\psi)$  maximum gain in scan direction  $\psi$ ,  $g_p$  limit obtained from projected dome area,  $g_{\min}$  minimum gain at  $\psi = 90^\circ$ . All curves normalized to feed array gain  $2\pi L/\lambda$  and  $\eta = 1$ .

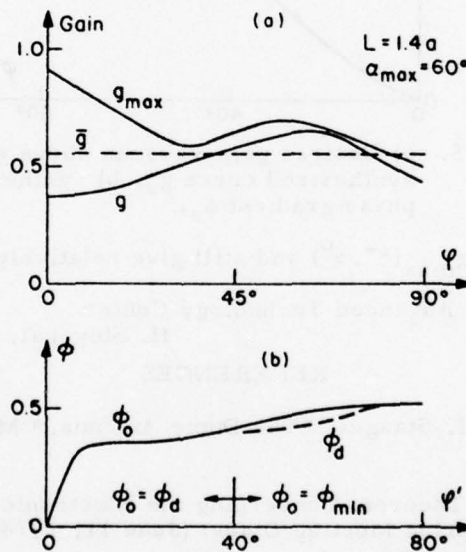


Fig. 4. a) Resultant gain-vs-scan curve  $g$  when dome phase gradient  $\phi_0$  is chosen to maximize gain at  $\psi = 67.5^\circ$ , b)  $\phi_0$  and actually desired gradient  $\phi_d$ . In sector  $40^\circ < \psi' < 80^\circ$   $\phi_0 < \phi_{\min}$ , and there  $\phi_0 = \phi_{\min}$  is used.

In conclusion, an upper bound on the gain-vs-scan average  $\bar{g}$  and the upper bound function  $g_{\max}(\psi)$  have been derived. These bounds are useful in posing realistic requirements on the dome antenna. Essentially, the feed array determines the area under the gain-vs-scan curve whereas the dome determines the maximum value of this curve,

once the scan angle limits have been set.

Two examples serve to illustrate the above points. Figure 5 shows gain-vs-scan curves obtained from a dome phase gradient chosen so as to maximize  $G(67.5^\circ)$ . The

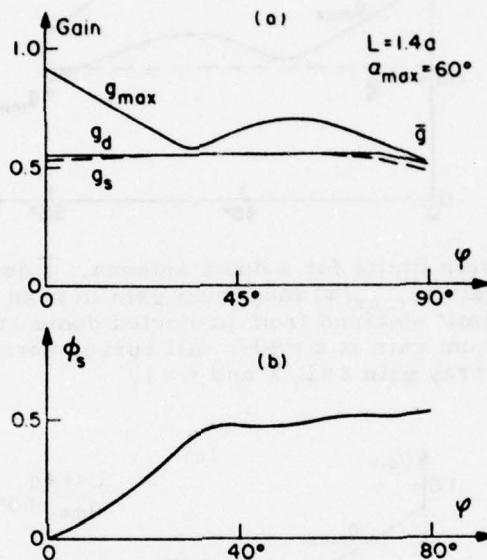


Fig. 5. a) Desired gain-vs-scan curve  $g_d$  and synthesized curve  $g_s$ , b) synthesized phase gradient  $\phi_s$ .

curves closely approach  $g_{\max}(67.5^\circ)$  and still give relatively high gain at  $\varphi = 90^\circ$ .

Ballistic Missile Defense Advanced Technology Center

DASG 60-76-C-0006

H. Steyskal, A. Hessel and J. Shmoys

#### REFERENCES

1. L. Schwartzman and J. Stangel, "The Dome Antenna," *Microwave Journal* (October 1975).
2. J. Stangel, "A Basic Theorem Concerning the Electronic Scanning Capabilities of Antennas," *URSI Spring Meeting Digest* (June 11, 1974).

## BLAZED DIFFRACTION GRATINGS FOR FREQUENCY SCANNED ANTENNAS

A. Hessel, J. Shmoys and S. T. Peng

A. Introduction

Frequency scanned traveling wave arrays are well-known and have received much attention in applications to line sources as well as to frequency-phase scanned phased array antennas. A different approach to frequency scanned phased arrays has been proposed by McDonnell-Douglas.<sup>1</sup> In this scheme, a Bragg angle blazed (or operated the Littrow condition) diffraction grating with one non-specular diffracted order is used as a reflect array, designed to frequency scan the  $n = -1$  grating order, while the specular beam is simultaneously suppressed.\* Alternatively, a transmission grating may be employed when blazed for Bragg angle of incidence in such a way that the specularly reflected ( $n = 0$ ) beam, as well as the transmitted ( $n = 0$ ) beam and the reflected ( $n = -1$ ) diffracted order are simultaneously suppressed. In that case, the transmitted  $n = -1$  grating order serves as the frequency scanned main beam. This study deals with the analysis and design method of a class of transmission gratings consisting of parallel bars of rectangular cross-section. We will show that it is possible to select grating parameters and a direction of incidence in the plane perpendicular to the grating elements (cf. Fig. 1) so as to achieve over 90% blazing efficiency with 30-40° scan range when the incident wave is taken to be polarized in the plane of incidence.

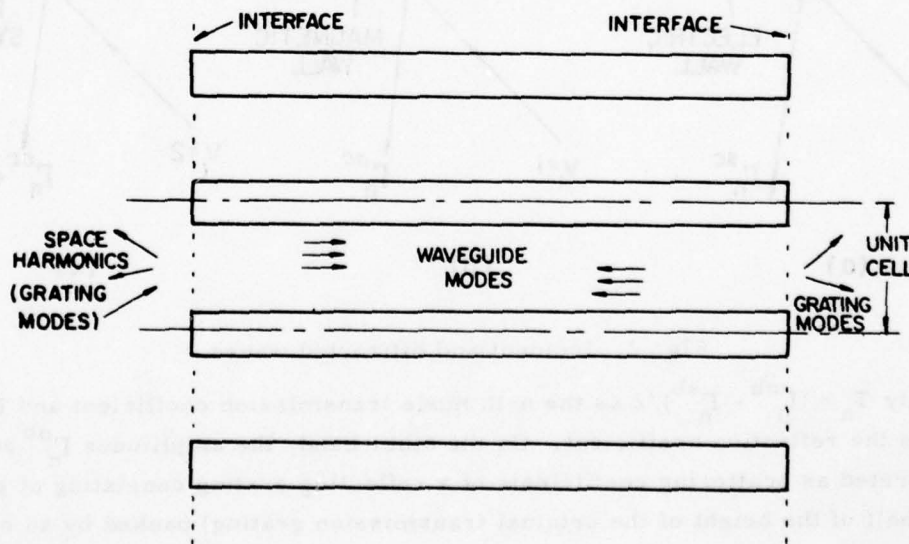


Fig. 1. Rectangular bar transmission grating.

\* In a blazed diffraction grating, the scattered power is concentrated in a particular spectral order. Blazing may be achieved by operating at Bragg angle of incidence ( $kd \sin \theta = \pi$ ). For an asymmetric grating profile, blazing can be attained both at Bragg angle and angles of incidence other than Bragg.

### B. Analysis

Let the transmission grating of Fig. 1 be illuminated in symmetrical fashion by two plane waves with electric field (component transverse to  $z$ -axis) amplitudes  $V_1$  and  $V_2$ . We will consider two special cases. If  $V_1 = -V_2$ , then the  $x$ - $y$  plane (mid-plane) can be replaced by a perfectly conducting sheet (electric wall) and we obtain a short circuit bisection of the structure (sb); if, on the other hand,  $V_1 = V_2$ , then the transverse to  $z$  magnetic field will vanish and we obtain the open-circuit bisection (ob), or magnetic wall.

The two cases are illustrated in Figs. 2(a) and 2(b) where the incident wave amplitudes as well as  $n$ -th space harmonic plane waves are indicated on both sides of the grating. Both sb and ob fields satisfy Maxwell's equations and all the boundary conditions imposed by the grating structure. Hence, the sum of the two fields also satisfies these conditions. For the sum we have, as shown in Fig. 2(c),  $V_1 = 2$  and  $V_2 = 0$ . Thus,

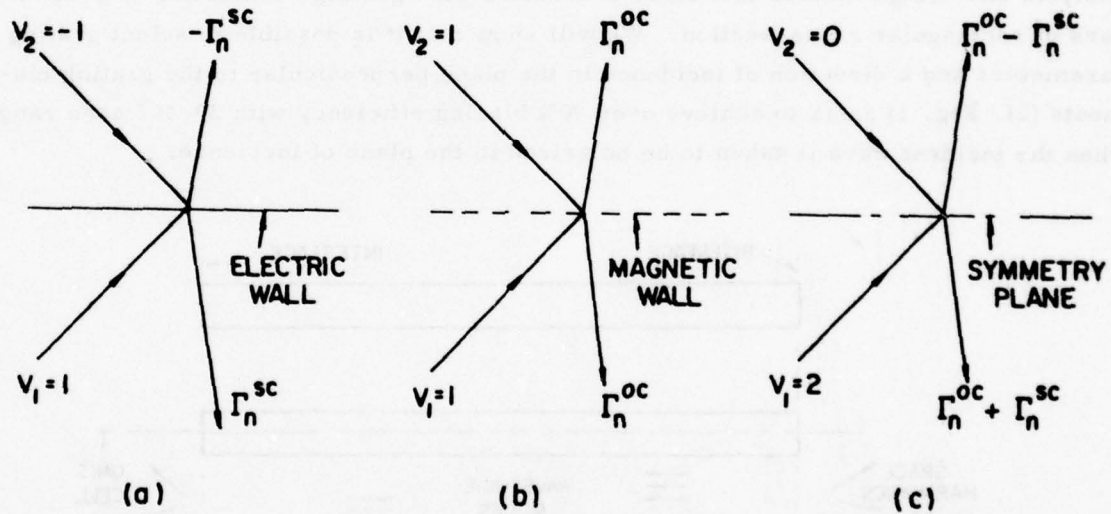


Fig. 2. Incident and diffracted waves.

we identify  $T_n = (\Gamma_n^{ob} - \Gamma_n^{sb})/2$  as the  $n$ -th mode transmission coefficient and  $R_n = (\Gamma_n^{ob} + \Gamma_n^{sb})/2$  as the reflection coefficient. On the other hand, the amplitudes  $\Gamma_n^{ob}$  and  $\Gamma_n^{sb}$  can be interpreted as scattering coefficients of a reflecting grating consisting of grooves of depth  $h$  (half of the height of the original transmission grating) backed by an electric or magnetic wall. The method of calculation of the scattering coefficients of a reflection grating was described previously<sup>2,3</sup> for the short-circuit case. The open-circuit bisection case constitutes a minor modification.

### C. Design

Transmission blazing of a symmetric grating imposes requirements on the scattering parameters of the grating, consequently on the reflection characteristics of the short-circuit bisection and open-circuit bisection of the structure. For a perfectly blazed transmission grating, one should operate in the propagation range of only two orders,  $n=0$  and  $n=-1$ , and one must have simultaneously

$$R_0 = T_0 = R_{-1} = 0 \quad (1)$$

where  $R_n$  and  $T_n$  are the reflection and transmission coefficients for the  $n$ -th order for the transmission grating. The first two of these conditions are satisfied if, and only if,

$$\Gamma_0^{ob} = \Gamma_0^{sb} = 0, \quad (2)$$

or, if both the mid-plane short-circuited reflection grating and the mid-plane open-circuited one are simultaneously blazed. Thus, we should look for a reflection grating which is blazed over a large range of depth, i.e., blaze depth vs. incidence angle characteristic, such as that of Fig. 3, which is nearly vertical. The cancellation of

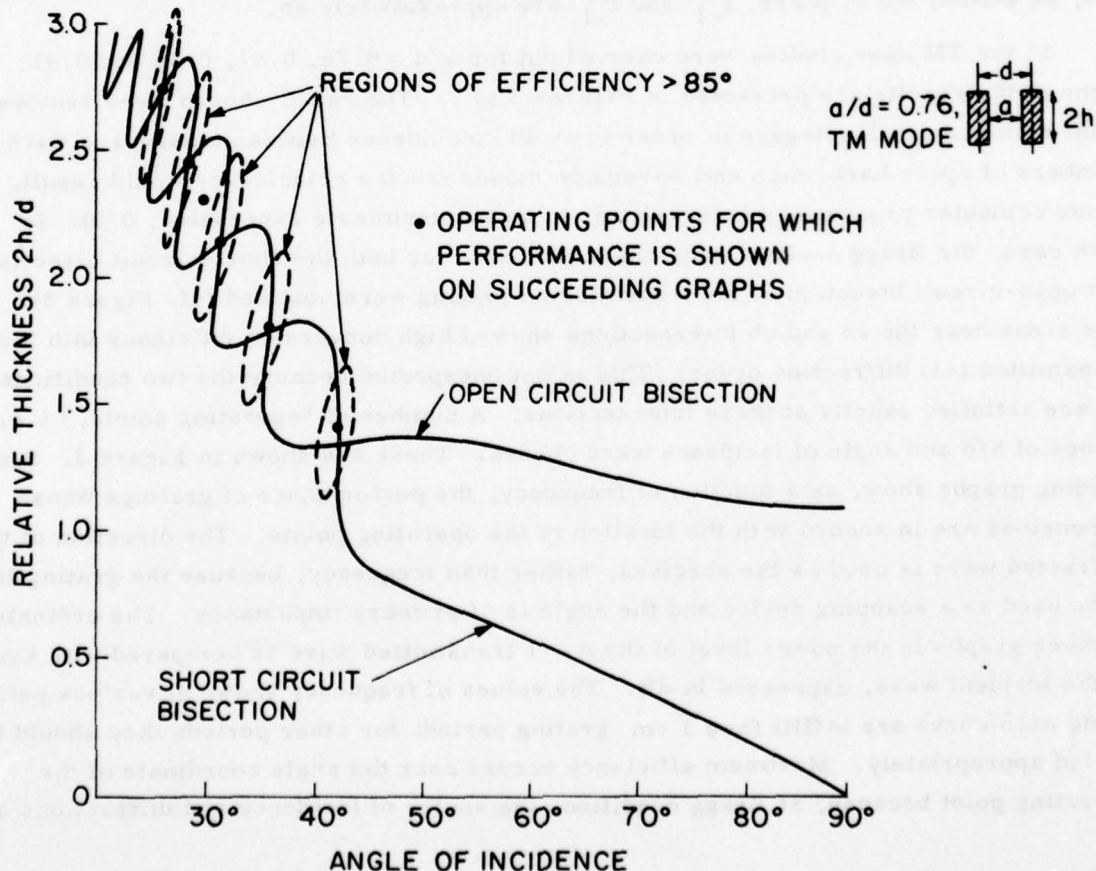


Fig. 3. Design curves for Bragg blazed transmission grating.

$n=0$  spectral orders both in reflection and transmission is predicated on the existence of the short-circuit and open-circuit bisections, i.e., on reflection symmetry and on the existence of the steep portions of the blazing curves. For  $a/d \geq 0.8$  such characteristics are indeed obtained and the mechanism responsible for this shape is the interference between two propagating parallel plate guide modes. The approximate cancellation of the  $n=-1$  grating mode is predicated on the  $\Gamma_{-1}^{ob}$  and  $\Gamma_{-1}^{sb}$  being approximately  $180^\circ$  out of phase at the blazing condition, and by power conservation  $|\Gamma_{-1}^{ob}| = |\Gamma_{-1}^{sb}|$ . The  $\Gamma_{-1}^{ob} \approx -\Gamma_{-1}^{sb}$  condition may indeed be realized at the Bragg blazing condition when the parallel plate guide free space interface discontinuity can be made small, i.e., for thin plates and TM polarization. In this case very little energy is back-scattered from the junction discontinuity and the dominant reflection is due to the wave entering the waveguide and reflected from the short-circuit or open-circuit respectively at the bottom of the groove. But the reflection coefficients for both of the propagating modes from the open-circuit are  $+1$  while from the short-circuit termination they are  $-1$ . The reflected parallel plate modes, after passing the discontinuity with a minor effect, re-radiate via the  $\Gamma_{-1}^{sb}$  or  $\Gamma_{-1}^{ob}$ . But, since the two reflections from sb and ob terminations are, as stated, out of phase,  $\Gamma_{-1}^{sb}$  and  $\Gamma_{-1}^{ob}$  are approximately so.

In the TM case studies were carried out for  $a/d = 0.76, 0.81, 0.86$  and  $0.91$ . Some of the results are presented in Figures 3 to 7. The ratios chosen were removed from ratios of small integers in order to avoid coincidence between transverse wave-numbers of space harmonics and waveguide modes (such a coincidence would result, in the computer program as it is written, in an indeterminate expression,  $0/0$ ). In each case, the Bragg angle blazing characteristics for both the short-circuit bisection and open-circuit bisection of the transmission grating were obtained (cf. Figure 3). The areas near the sb and ob intersections showed high conversion efficiency into the transmitted ( $-1$ ) diffraction order. This is not unexpected because the two conditions (1) are satisfied exactly at these intersections. A number of "operating points," i.e., values of  $h/d$  and angle of incidence were chosen. These are shown in Figure 3. Succeeding graphs show, as a function of frequency, the performance of gratings whose dimensions are in accord with the location of the operating points. The direction of the diffracted wave is used as the abscissa, rather than frequency, because the grating is to be used as a scanning device and the angle is of primary importance. The ordinate in these graphs is the power level of the  $n=-1$  transmitted wave as compared with that of the incident wave, expressed in dB. The values of frequency shown at various points along each curve are in GHz for a 1 cm grating period; for other periods they should be scaled appropriately. Maximum efficiency occurs near the angle coordinate of the operating point because, at Bragg condition, the angles of incidence and diffractions are

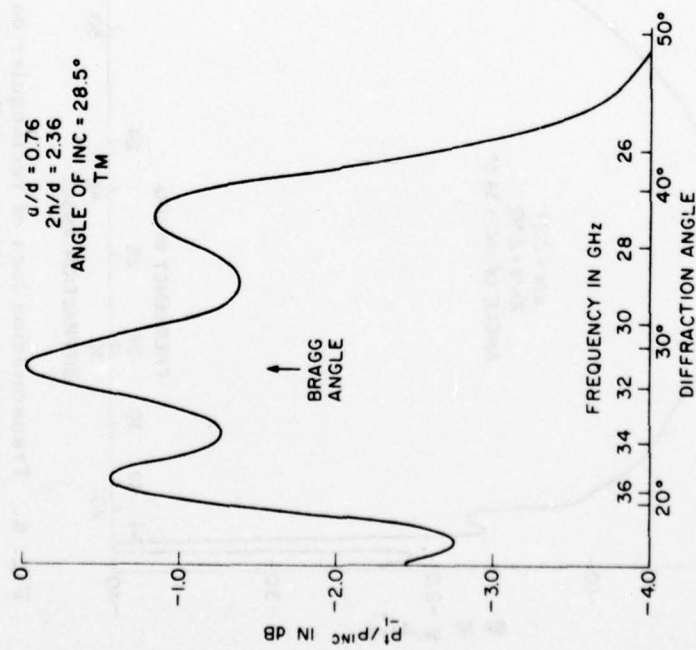


Fig. 4. Transmission loss of rectangular bar transmission grating vs. diffraction angle (or frequency).

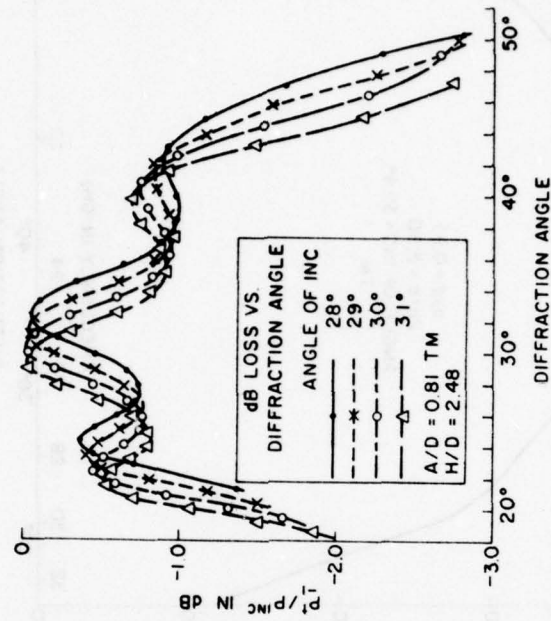


Fig. 5. Transmission loss of rectangular bar transmission grating vs. diffraction angle (or frequency).

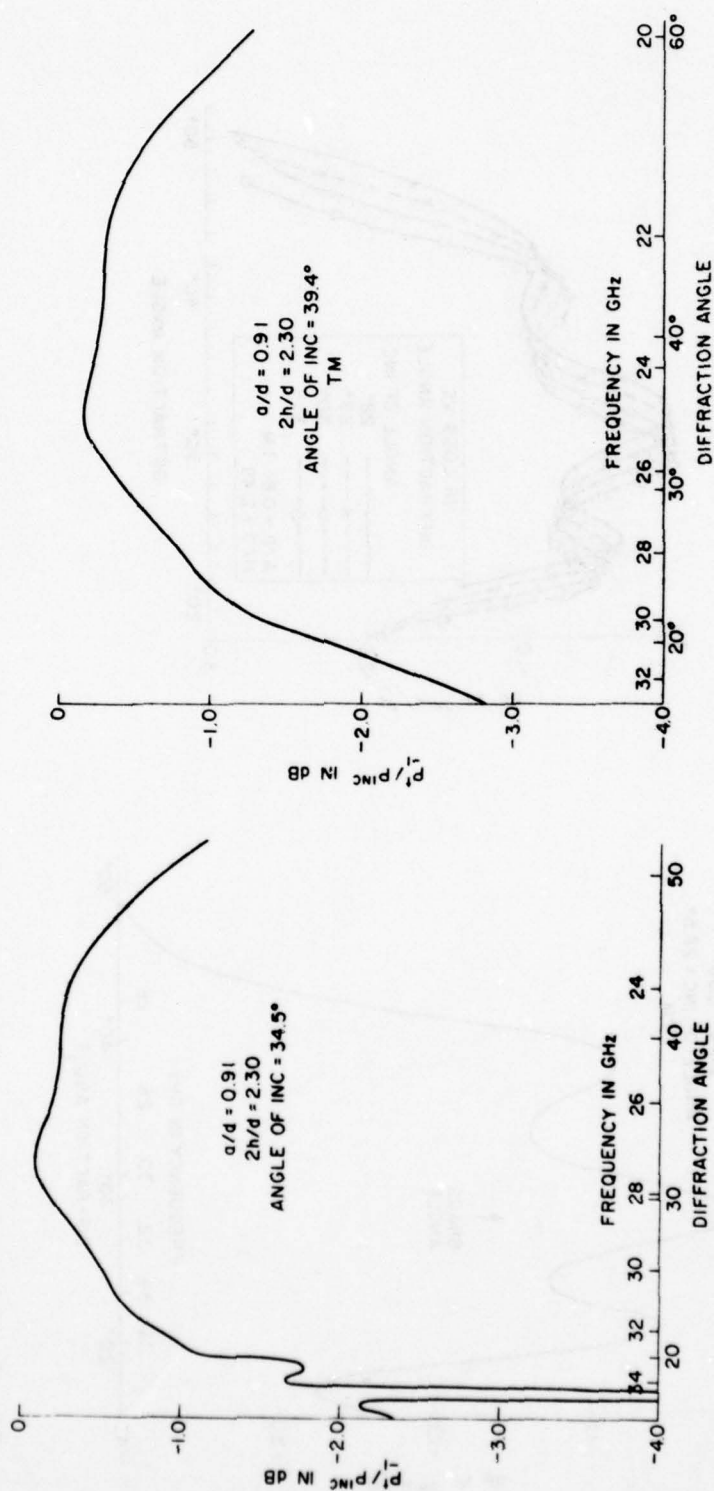


Fig. 6. Transmission loss of rectangular bar transmission grating vs. diffraction angle (or frequency).

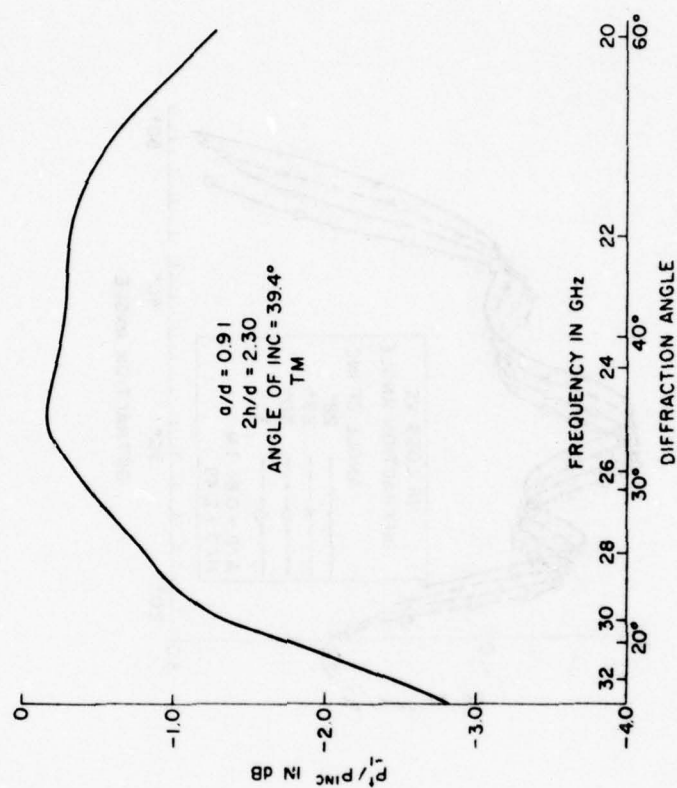


Fig. 7. Transmission loss of rectangular bar transmission grating vs. diffraction angle (or frequency).

numerically equal. Subsidiary peaks appear to be related to neighboring ovals of high efficiency (cf. Figure 3). Figure 4 shows typical performance of a grating designed according to one of the intersections of short-circuit and open-circuit curves of Figure 3. The effect of a small change in the angle of incidence is shown, for a grating with somewhat thinner bars, in Figure 5. Efficiency over the pass-band changes little, but the angular range is displaced. The thinner the grating, the broader the angular range of its usefulness. This, again, appears to be related to the spacing of the ovals and to the fact that the thinner the wall the smaller the effect of the junction discontinuity. The resulting improved performance is shown in Figures 6 and 7. In these cases the bar thickness is 9% of the grating period, and remaining parameters correspond to two intersections of open- and short-circuit blazing curves. Less than 1 dB conversion loss (not considering ohmic losses) can be achieved over an angular scan range of more than  $30^\circ$ .

Limited studies were carried out for polarization parallel to grating elements (TE case), and, in the TM case, for a grating of rectangular dielectric bars. In both these cases very high conversion efficiency was obtained for gratings designed using the procedure described above, but the bandwidth and the useful angular scan range was found to be much more restricted.

Subcontract from McDonnell Douglas Astronautics Co.

on Contract DASG60-77-C-0067, U.S. Army

Ballistic Missile Defense Advanced Technology Center A. Hessel, J. Shmoys and S. T. Peng

#### REFERENCES

1. R.A. Profet, R.A. Willett and T.O. Tobin, "Dispersively Scanned Radar," Final Report, Vol. 1, McDonnell Douglas Astronautics Company, Calif., Report MDC G7532 (August 1978).
2. A. Hessel and J. Shmoys, "Study of Blazed Gratings," Progress Report No. 36 to JSTAC, Polytech. Inst. of New York, Report R-452.36-71, pp. 88-91 (1971).
3. A. Hessel and J. Shmoys, "Bragg-angle Blazing of Diffraction Gratings," J. Opt. Soc., Vol. 65, No. 4, pp. 380-384 (April 1975).

## PROPERTIES OF THE SHADOW CAST BY A HALF-SCREEN WHEN ILLUMINATED BY A GAUSSIAN BEAM

A. C. Green, H. L. Bertoni and L. B. Felsen

A. Introduction

Optical components such as irises, occulting knife edges and stray light baffles are designed for their shadow producing properties. For simple fields that appear to diverge from a point in the three dimensional case, a line in the two dimensional case, or a pair of perpendicular lines in the astigmatic three dimensional case, the location of the shadow boundary and the nature of the fields away from the shadow boundary can be found using the geometrical theory of diffraction (GTD). Such simple divergent fields can be described in terms of rays along which the fields propagate. The rays that graze the edge of the shadowing obstacle, which we call critical rays, define the shadow boundary. Treating the critical rays as if they were reflected defines the shadow boundary of the reflected field. Finally, the critical rays can be viewed as generating the diffracted rays that are the source of the fields in the shadow region. The simple distinction between incident, reflected and diffracted fields is not valid within transition regions centered about the shadow boundaries. Within these regions, a uniform theory of diffraction must be employed, which gives the continuous variation of the field across the shadow boundary.

Optical fields are however commonly in the form of bounded beams, which cannot be described in terms of real rays.<sup>1,2</sup> In the case of Gaussian beams, the fields can be described in terms of rays emanating from a source point having complex coordinates. These rays must be viewed as traveling through complex space and intersecting real space at a single point. Because each of these complex rays illuminates a single point in real space, the critical rays that "graze" the shadowing obstacle cannot define the shadow boundaries of the regions illuminated by the incident or reflected fields.

The complex ray description of a Gaussian beam was previously used to find the diffracted fields generated by an iris.<sup>3</sup> However, the shadow boundaries were located on the basis of an ad hoc assumption, rather than a rigorous analysis, and no consideration was given to the shape of the transition regions, or to the field variation within them. For the case of an inhomogeneous plane wave incident on a half-plane, the location of the shadow boundary was found, as well as the shape of the transition regions and the variation of the fields within them.<sup>4,5</sup>

In this study, we examine the location of the shadow boundaries that result when a two-dimensional or ribbon beam having Gaussian profile is incident on a perfectly conducting half-plane, as shown in Figure 1. The concept of a shadow boundary is

associated with the high frequency asymptotic approximation of the total field, which separates the field into incident, reflected and diffracted components. We, therefore, start with an exact integral representation for the total field, and approximate the integral asymptotically.

Except when the beam axis passes through the edge, the shadow boundary of the incident (reflected) beam field is shown to be a Stokes line of the integral representation. As a result, the incident (reflected) field is of exponentially smaller order than the diffracted field. Since the fields are in fact continuous across the shadow boundary, taking the incident (reflected) field to exist on one side, but not on the other, introduces an error that is on the order of the incident (reflected) field, and hence of exponentially small order compared to the diffracted field. Unlike the case of a real line source, or when the screen edge lies on the beam axis, the diffracted fields produced by a beam are finite and continuous across the shadow boundaries. Near the screen edge, the shadow boundaries for the incident beam also coincide with those of an inhomogeneous plane wave, although they deviate far from the edge. As the edge approaches the beam axis, the shadow boundary approaches that of a real line source located at the center of the beam throat.

The asymptotic approximation that allows separation of incident, reflected and diffracted field components is valid outside of some transition regions. The size and shape of these regions is dependent on the accuracy stipulated for the approximation, and on the beam width and edge location. Within the transition regions, a uniform asymptotic expression must be used for the field. In the case of a real line source fields, the transition regions extend to infinity and are centered about the shadow boundaries. By contrast, it is shown here that the transition regions for the incident beam are not centered on the shadow boundaries, except when the beam axis passes through the screen edge. Moreover, if the edge is located far enough from the beam axis, the transition regions do not extend to infinity, but are instead bounded, as is the case when an inhomogeneous plane wave is incident. When the edge is near enough to the beam axis, the transition regions extend to infinity. If the edge is located in the far field region of the beam, the shadow boundaries and transition regions approach those of a line source centered at the beam throat.

#### B. Formal Solution for the Diffracted Fields

It has previously been shown that the free-space Green's function, which is an exact solution to Maxwell's equations, gives the fields of a Gaussian beam when complex values are assigned to the source coordinates.<sup>1,2</sup> Starting with the point-source Green's function, a Gaussian beam with circular cross-section is obtained. Using the

line-source Green's function gives a ribbon beam whose profile is Gaussian. The advantage of using the Green's function to obtain a beam field is that expressions derived for the propagation and diffraction of point-source or line-source fields can be used to study the propagation and diffraction of beams.

For this study we consider a two-dimensional or ribbon beam propagating parallel to the  $z$ -axis with no variation along  $y$ , as depicted in Figure 1. If the electric field is polarized along  $y$ , the beam field  $E_1(\underline{\rho}; \underline{\rho}')$  are obtained from the free-space Green's function for the current line source

$$\underline{J} = \underline{y}_0 \frac{1}{i\omega\mu_0} \delta(x-x') \delta(z-z') e^{-i\omega t} \quad (1)$$

with complex source coordinates

$$\left. \begin{aligned} x' &= -x_0 \\ z' &= -z_0 + ib \end{aligned} \right\} \quad (2)$$

In Eq. (1),  $\delta(x-x')$  is the delta function,  $\omega$  the radian frequency and  $\mu_0$  the permeability of free-space. The electric field generated by the source is

$$E_1(\underline{\rho}; \underline{\rho}') = \frac{i}{4} H_0^{(1)}(kR); \quad R = \sqrt{(x+x_0)^2 + (z+z_0-ib)^2} \quad (3)$$

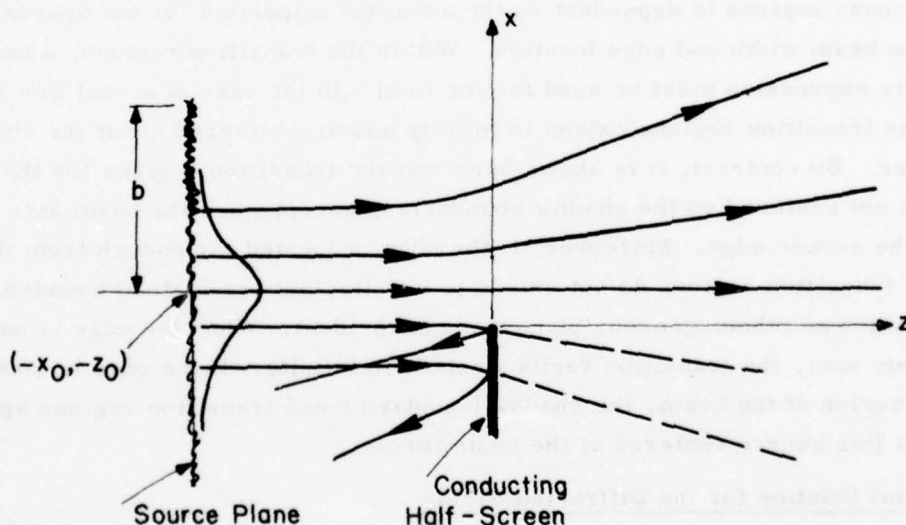


Fig. 1. Gaussian beam equivalent to the field radiated by a line source at the complex location  $(-x, -z_0 + ib)$ . The beam is indicated by phase paths along which the exponential amplitude of the field is constant.

where  $H_0^{(1)}(kR)$  is the zero order Hankel function of the first kind, and  $k = \omega/C$  is free-space wavenumber. The complex displacement  $R$  has branch point singularities at  $x = -x_0 \pm b$ ,  $z = -z_0$  where  $R = 0$ . In order for  $R$ , and hence  $E_i(\underline{\rho}; \underline{\rho}')$  to be single valued, it is necessary to connect the branch points with a cut, such as the one shown in Figure 1. The field is discontinuous across the cut, which requires the presence of sources. Viewed another way, by placing the proper source distribution along the cut, the radiated fields will be those of Equation (3). For the choice of cut in Fig. 1,  $\text{Re}(R) \geq 0$  and Eq. (3) represents an outgoing wave for  $\rho \rightarrow \infty$ .

Far from the branch points,  $k|R| \gg 1$  and the Hankel function may be replaced by its asymptotic approximation so that

$$E_i(\underline{\rho}; \underline{\rho}') \sim \frac{\exp[i(kR + \pi/4)]}{2\sqrt{2\pi kR}} \quad (4)$$

In the paraxial region  $(x+x_0)^2 \ll (z+z_0)^2 + b^2$  and we may approximate  $R$  in Equation (4). For  $z > -z_0$  we obtain

$$E_i(\underline{\rho}; \underline{\rho}') \sim \frac{\exp[kb + ik(z-z_0) + i\pi/4]}{2\sqrt{2\pi} \sqrt{k(z+z_0 - ib)}} \exp\left\{-\frac{(x-x_0)^2}{w^2} \left(1 - i\frac{z+z_0}{b}\right)\right\} \quad (5)$$

where

$$w = \sqrt{2 \frac{(z+z_0)^2 + b^2}{kb}} \quad (6)$$

Expression (5) is the field of a Gaussian beam whose  $1/e$  half-width is  $w$ . For  $z+z_0 \rightarrow 0$  the half-width is seen to be  $\sqrt{2b/k}$  so that the parameter  $b$  represents the classical distance to the far field. In the Fraunhofer or far field region where  $z+z_0 \gg b$ , the beam diverges and appears to come from a line source located at the center  $(-x_0, -z_0)$  of the beam throat.

The field Eq. (3) is assumed to be incident from the left on a conducting half-screen occupying the half-plane  $z=0$ ,  $x \leq 0$ , as shown in Figure 1. An expression for the total field radiated by a real line source in the presence of the screen has previously been derived.<sup>6,7</sup> Letting the source have complex location, we obtain the total field for the beam case as

$$E(\underline{\rho}; \underline{\rho}') = \frac{\exp(ikR^+)}{2\pi} I^+ - \frac{\exp(ikR^-)}{2\pi} I^- \quad (7)$$

\* This result is obtained from Ref. [6] with the substitution  $\mu = t \exp(i\pi/4)$ .

where

$$I^{\pm} = \int_{W^{\pm}}^{\infty} \frac{e^{-t^2}}{\sqrt{t^2 - i2kR^{\pm}}} dt \quad (8)$$

In Eqs. (7) and (8),  $R^+$  and  $R^-$  are the complex distances to the observation point from the source at  $(-x_0, -z_0 + ib)$  and from the image source at  $(-x_0, z_0 - ib)$ , respectively. These distances are given by

$$R^{\pm} = \sqrt{(x+x_0)^2 + [z \mp (-z_0 + ib)]^2} \quad (9)$$

where the root is chosen such that  $\text{Re}(R^{\pm}) > 0$ . The end points  $W^{\pm}$  of the integration are defined below.

Let  $(\rho, \theta)$  with  $-\pi < \theta < \pi$  be the polar coordinate of the observation point and  $(\rho', \theta')$  be the polar coordinates of the source given by

$$\left. \begin{aligned} \theta' &= \tan^{-1} \left[ \frac{x_0}{z_0 - ib} \right] \\ \rho' &= \sqrt{x_0^2 + (z_0 - ib)^2} \end{aligned} \right\} \quad (10)$$

with  $\text{Re}(\rho') > 0$  and  $0 < \text{Re}(\theta') < \pi$ . The angles  $\theta$  and  $\theta'$  are indicated in Fig. 2(a) for a real line source ( $b = 0$ ).

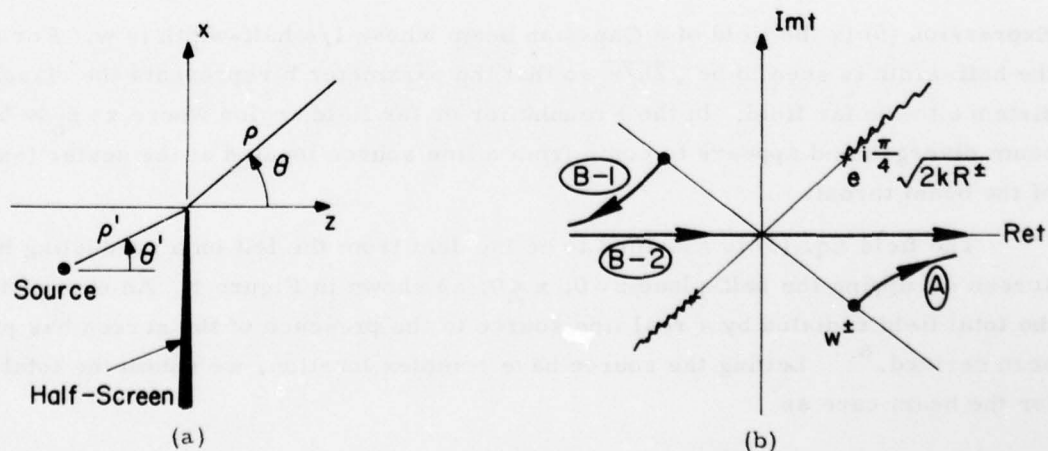


Fig. 2. (a) Polar coordinate system showing source coordinates for a real line source. (b) Integration path and branch points of the integral representation for the total field.

The end points  $W^\pm$  of the integrations in Eq. (8) are given in terms of the polar coordinates as

$$W^\pm = 2e^{-i\pi/4} \sqrt{\frac{k\rho\rho'}{\rho + \rho' + R^\pm}} \begin{cases} -\sin \frac{1}{2}(\theta - \theta') \\ \cos \frac{1}{2}(\theta + \theta') \end{cases} \quad (11a)$$

or

$$W^\pm = -e^{-i\pi/4} \sqrt{k(\rho + \rho' - R^\pm)} \quad (11b)$$

For a complex source,  $W^\pm = 0$  only at the screen edge  $\rho = 0$ , so that  $W^\pm$  is single valued in the domain  $-\pi/2 < \theta < 3\pi/2$ . The choice of square root in Eqs. (11a) or (11b) is made so that  $\text{Re } W^\pm > 0$  for  $\rho \rightarrow \infty$  with  $\theta \rightarrow -\pi/2$ , since as argued in the next section only the diffracted field reaches the far side of the screen.

Steepest descent paths for the integrations in Eq. (7) are indicated in Fig. 2(b) for the case (A) when  $\text{Re}(W^\pm) > 0$  and case (B) when  $\text{Re}(W^\pm) < 0$ . The branch points  $\pm e^{i\pi/4} \sqrt{2kR^\pm}$  and hill regions of the integrands are also indicated in the figure. For later use, we may express  $R^\pm$  in terms of the polar coordinates as

$$R^\pm = \sqrt{\rho^2 \pm 2\rho\rho' \cos(\theta \mp \theta') + (\rho')^2} \quad (12)$$

### C. Asymptotic Approximation for the Fields and the Shadow Boundary

In the geometrical theory of diffraction (GTD), the total field  $E(\underline{\rho}, \underline{\rho}')$  is separated into a geometrical optics component  $E_{GO}(\underline{\rho}, \underline{\rho}')$  and a diffraction component  $E_D(\underline{\rho}, \underline{\rho}')$  as

$$E(\underline{\rho}; \underline{\rho}') = E_D(\underline{\rho}; \underline{\rho}') + E_{GO}(f, f\rho') \quad (13)$$

The geometrical optics component consists of the incident and reflected fields, which exist in regions of space bordered by shadow boundaries. The diffraction component includes the remainder of the field.

Concepts such as the shadow boundary that are basic to GTD arise out of the asymptotic approximation for the field Equation (7). The asymptotic approximation for the integral in Eq. (8) may be obtained by taking the integration path to be the steepest descent path away from the end point  $W^\pm$ . For  $\text{Re}(W^\pm) > 0$  the steepest descent path runs to  $t = +\infty$  asymptotic to the real  $t$  axis, as shown by path (A) in Figure 2(b). When  $\text{Re}(W^\pm) < 0$ , the steepest descent path runs asymptotic to the negative real  $t$  axis.

In this case it is necessary to return to the end point  $t = +\infty$  via the saddle point at  $t = 0$ , as shown by path (B) in Figure 2(b). Thus the saddle point contributes to Eq. (8) for  $\text{Re}(W^\pm) < 0$ , but not for  $\text{Re}(W^\pm) > 0$ .

Let  $\theta_{si}(\rho)$  be the locus of points for which  $\text{Re}(W^+) = 0$ , and  $\theta_{sr}(\rho)$  be the locus points for which  $\text{Re}(W^-) = 0$ . For  $\theta > \theta_{si}(\rho)$ ,  $\text{Re}(W^+) < 0$  and the saddle point contributes to  $I^+$ . Similarly, for  $\theta > \theta_{sr}(\rho)$ ,  $\text{Re}(W^-) < 0$  and the saddle point contributes to  $I^-$ . The saddle point contributions can be expressed in terms of Hankel functions,<sup>6</sup> and are found to allow their interpretation as being the geometrical optics contribution to the field. Thus,

$$E_{GO}(\rho; \rho') = u[\theta - \theta_{si}(\rho)] H_0^{(1)}(kR^+) - u[\theta - \theta_{sr}(\rho)] H_0^{(1)}(kR^-), \quad (14)$$

where the first term gives the incident field and the second term gives the reflected field. The locii of points  $\theta_{si}(\rho)$  and  $\theta_{sr}(\rho)$  are seen to be the shadow boundaries of the incident and reflected fields, respectively.

In the case of a real line source ( $b = 0$ ),  $\theta'$  and  $R^\pm$  are real and hence from Eq. (11)  $\text{Re}(W^+) = 0$  when  $\sin \frac{1}{2}(\theta - \theta') = 0$ , while  $\text{Re}(W^-) = 0$  when  $\cos \frac{1}{2}(\theta + \theta') = 0$ . Thus  $\theta_{si} = \theta'$  and  $\theta_{sr} = \pi - \theta_{si}$ , so that the two shadow boundaries are straight lines originating at the edge, and are symmetrically located about the x-axis. The shadow boundaries coincide with the incident and reflected rays that pass through the edge.

For a line source located at a complex point,  $\theta_o$  and  $R^\pm$  are complex so that  $\theta_{si}(\rho)$  and  $\theta_{sr}(\rho)$  do not have a simple geometrical interpretation, although we still have that  $\theta_{sr}(\rho) = \pi - \theta_{si}(\rho)$ . In order to locate the shadow boundary  $\theta_{si}(\rho)$ , note that if  $\text{Re}(W^+) = 0$ , then  $\arg(W^+)^2 = \pi$ . With the help of Eq. (11b) this last condition is seen to require that  $\text{Re}(\rho + \rho' - R^+) = 0$  with  $\text{Im}(\rho + \rho' - R^+) < 0$ . Simple expressions for  $\theta_{si}(\rho)$  are obtained in the limits  $\rho \ll |\rho'|$  and  $\rho \gg |\rho'|$ .

For  $\rho \ll |\rho'|$  we approximate  $R^+$  From Eq. (12) as

$$R^+ \approx \rho' + \rho \cos(\theta - \theta') \quad (15)$$

and obtain for the shadow boundary  $\theta_{si}(0)$  near the edge the expression

$$\tan[\theta_{si}(0) - u] = \sinh v \quad (16)$$

where

$$\left. \begin{aligned} u &= \text{Re}(\theta') \\ v &= \text{Im}(\theta') \end{aligned} \right\} \quad (17)$$

The location Eq. (16) is the same as was previously found for an inhomogeneous plane wave incident on the edge when the wave vector of the plane wave makes the (complex) angle  $\theta'$  with the  $z'$ -axis.<sup>5</sup> In the vicinity of the edge, the incident field radiated by the complex line source is locally that of an inhomogeneous plane wave whose wave vector  $\nabla(kR)$  makes the angle  $\theta'$  with the  $z$ -axis, thus accounting for the location of the shadow boundary in the vicinity of the edge.

Assuming the edge to be in the paraxial region of the incident beam ( $x_0 \ll \sqrt{z_0^2 + b^2}$ ), it is seen from Eq. (10) that  $|\theta'| \ll 1$  and hence

$$\left. \begin{aligned} u &\approx \operatorname{Re} \left( \frac{x_0}{z_0 - ib} \right) = \frac{x_0 z_0}{z_0^2 + b^2} \\ v &\approx \operatorname{Im} \left( \frac{x_0}{z_0 - ib} \right) = \frac{x_0 b}{z_0^2 + b^2} \end{aligned} \right\} \quad (18)$$

For  $|v| \ll 1$ ,  $\tan^{-1}(\sinh v) \approx v$  so that from Eq. (16)

$$\theta_{si}(0) \approx u + v = \frac{x_0(z_0 + b)}{z_0^2 + b^2}, \quad (19)$$

which is the angle the shadow boundary makes with the  $z$ -axis for  $\rho \ll |\rho'|$ . The shadow boundary is indicated in Fig. 3 for the case  $x_0 > 0$ , which occurs when the edge blocks the beam axis. When the edge lies on the beam axis, the shadow boundary coincides with the beam axis.

For  $\rho \gg |\rho'|$ , i.e., far from the edge, we may approximate  $R^+$  from Eq. (12) as

$$R^+ \approx \rho + \rho' \cos(\theta - \theta')$$

and obtain for the shadow boundary  $\theta_{si}(\infty)$  the expression

$$\tan[\theta_{si}(\infty) - u] = \sinh v \frac{\operatorname{Re}(\rho')}{|\rho'| + \operatorname{Im}(\rho') \cosh v} \quad (21)$$

The location of the shadow boundary  $\theta_{si}(\infty)$  for  $\rho \gg |\rho'|$  depends strongly on the location of the edge within the beam through the factor multiplying  $\sinh v$  in Equation (21).

In the paraxial region  $x_0^2 \ll z_0^2 + b^2$ , and From Eq. (10) we have

$$\rho' \approx z_0 \left( 1 + \frac{1}{2} \frac{x_0^2}{z_0^2 + b^2} \right) - ib \left( 1 - \frac{1}{2} \frac{x_0^2}{z_0^2 + b^2} \right) \quad (22)$$

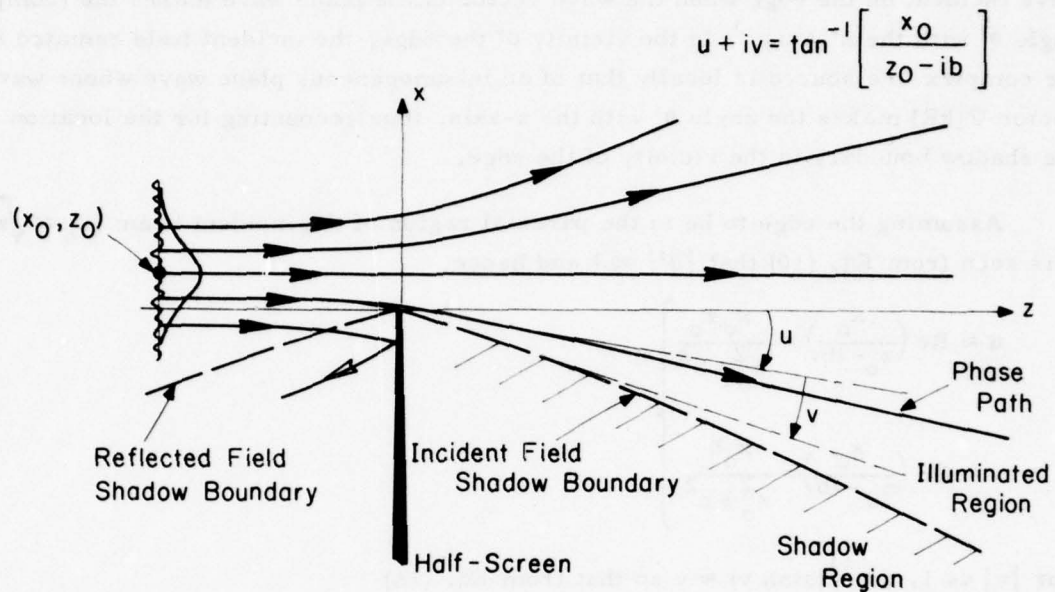


Fig. 3. Shadow boundaries of the incident and reflected fields.

so that  $\text{Im}(\rho') < 0$ . Because

$$\text{Re}(\rho') + |\text{Im}(\rho')| \cosh v \geq \text{Re}(\rho') + |\text{Im}(\rho')| \geq |\rho'| \quad (23)$$

it is seen that for  $\text{Im}(\rho') < 0$ ,

$$\text{Re}(\rho') > |\rho'| + \text{Im}(\rho') \cosh v \quad (24)$$

Thus the factor multiplying  $\sinh v$  in Eq. (21) will be greater than unity. Comparing Eqs. (16) and (21) we see that

$$\theta_{si}(\infty) > \theta_{si}(0) \quad (25)$$

when the edge is located in the paraxial region. As a result, the shadow boundary turns away from the  $z$ -axis for  $\rho \rightarrow \infty$ , as is shown in Figure 3. Should the denominator of Eq. (21) vanish, then  $\theta_{si}(\infty) - u = \pi/2$  and shadow boundary will run to  $\rho \rightarrow \infty$  at angle  $u$  from the  $x$ -axis. Consistent with paraxial approximation Eq. (22), the denominator vanishes for  $|x_0| \approx z_0 \ll b$ , i.e., when the edge is located in the near field along a line at  $45^\circ$  from  $z$ -axis. For edge locations above this  $45^\circ$  line, the shadow boundary turns further behind the  $x$ -axis into the region  $z < 0$ .

The location of the shadow boundary  $\theta_{si}(\rho)$  is shown in Fig. 3 for the case  $x_0 > 0$  so that edge blocks the main part of the beam. For  $x_0 < 0$ , the shadow boundary is located symmetrically below the  $z$ -axis. For comparison, we have also drawn in Fig. 3 the curve passing through edge along which  $\text{Im } \rho'$  has constant value. This curve represents a contour along which the exponential amplitude of the incident beam is constant. When the edge is located in the paraxial region of the beam, the curve has slope

$$\frac{dx}{dy} = \left\{ \begin{array}{ll} \frac{x_0 z_0}{z_0^2 + b^2} = u & (\rho \rightarrow 0) \\ \frac{x_0}{z_0^2 + b^2} & (\rho \rightarrow \infty) \end{array} \right\} \quad (26)$$

The slope for  $\rho \rightarrow \infty$  has magnitude smaller than  $|u+v|$ , as indicated in Figure 3.

It is seen that the shadow boundary curves into a region where the incident beam field is exponentially small compared to its value at the edge. Since the diffracted field  $E_D$  everywhere has exponential order equal to the order of the incident beam at the edge, along the shadow boundary  $E_D$  is exponentially larger than the incident field  $E_{GO}$ . Thus, as the shadow boundary is crossed, a discontinuity in the field is obtained whose magnitude is equal to  $|E_{GO}|$ , which is exponentially smaller than the diffracted field there. In the literature on asymptotic analysis, this behavior is referred to as the Stokes phenomena and the shadow boundary is referred to as a Stokes line.

The properties of the shadow boundary cited above are in accordance with those found for an inhomogeneous plane wave blocked by an edge. Near the edge the shadow boundary has the same location as would be obtained for the inhomogeneous plane wave that approximates the incident beam field in the vicinity of the edge. However, far from the edge the shadow boundary is shifted, indicating that its location cannot be predicted solely from the properties of the field in the vicinity of the edge, but will also depend on the properties of the field far from the edge.

#### D. Diffracted Field and the Transition Region

The diffracted field  $E_D(\underline{\rho}; \rho')$  is given by the integration in Eq. (8) over the path segment from the end point  $W^\pm$  to  $|t| \rightarrow \infty$ . These segments are indicated in Fig. 2(b) as (A) for  $\text{Re } W^\pm > 0$  and as (B-1) for  $\text{Re } W^\pm < 0$ . For  $|w| \gg 1$ , these integrations can be approximated asymptotically by the method of steepest descent. Retaining the first two non-vanishing terms in the asymptotic approximation gives

$$E_D(\rho; \rho') = E_i(0; \rho') \frac{e^{i(k\rho + \pi/4)}}{\sqrt{2\pi k\rho}} \left\{ \frac{1 - \Delta^+}{\sqrt{1 - \cos(\theta - \theta')}} - \frac{1 - \Delta^-}{\sqrt{1 + \cos(\theta + \theta')}} \right\} \quad (27)$$

where

$$\Delta^\pm = \frac{i(\rho + \rho')}{2k\rho\rho'[1 \mp \cos(\theta + \theta')]} \quad (28)$$

In Eq. (27),  $E_i(0, \rho')$  is the field incident on the edge, as given by Eqs. (4) or (5) with  $R = \rho'$ .

Neglecting the terms  $\Delta^\pm$  in Eq. (27) yields the geometrical theory of diffraction (GTD) approximation  $E_{GTD}$  for the diffracted field. The expression for  $E_{GTD}$  is the same as was previously obtained for an inhomogeneous plane wave propagating at a complex  $\theta'$  and incident on the edge. The terms  $\Delta^\pm$  represent corrections to  $E_{GTD}$ . Thus, the accuracy of  $E_{GTD}$  is determined by  $|\Delta^\pm|$ .

For a real line source  $\theta'$  is real and the field is singular across the shadow boundaries. Because  $\theta'$  is complex for a beam, the field is never singular. However  $|\Delta^+|$  may be significant compared to unity in the vicinity of the shadow boundary for the incident field. Similarly  $|\Delta^-|$  may be significant in the vicinity of the shadow boundary of the reflected field. In those regions of space where  $\Delta^\pm$  does have a significant magnitude, the asymptotic approximation cannot be used and a more accurate evaluation of the integrals  $I^\pm$  is required. By analogy to the case of a real line source, these regions in space are called transition regions where the GTD expressions for the diffracted field are not valid.

The transition regions can be defined as the locations in space where the relative error made in using  $E_{GTD}$  for  $E_D$  has magnitude less than some value  $1/A$ , where  $A$  is a large number. Then the transition regions are defined by

$$|\Delta^\pm| \geq 1/A \quad (29)$$

The boundary of the transition region is obtained from the equality in Equation (29).

Using Eq. (28) in Eq. (29), we obtain for the boundary of the incident field's transition region the equation

$$|\rho + \rho'| A = 2k\rho |\rho'| |1 - \cos(\theta - \theta')| \quad (30)$$

Substituting  $\theta' = u + iv$  and squaring both sides gives

$$(\rho^2 + 2\rho \operatorname{Re} \rho' + |\rho'|^2) = \frac{4k^2 |\rho'|^2}{A^2} \rho^2 \left\{ [1 - \cos(\theta - u) \cosh v]^2 + \sin^2(\theta - u) \sinh^2 v \right\}, \quad (31)$$

whose solution for  $\rho$  as a function of  $\theta$  gives the boundary of the transition region. It is seen that the transition region is symmetric about the radius  $\theta = u$  extending from the edge. To find the maximum distance that the transition region extends away from the edge, set  $\theta = u$  in Eq. (31) and solve the resulting quadratic for  $\rho$ . Since  $\operatorname{Re} \rho' > 0$  the quadratic will have a real solution with  $\rho > 0$  only if

$$\frac{4k^2 |\rho'|^2}{A^2} [1 - \cos v]^2 > 1. \quad (32)$$

If the inequality Eq. (32) is not satisfied, no positive real solution will exist for  $\rho$ , indicating that the transition region extend to  $\rho = \infty$ , as happens for a real line source. The case of the transition region extending to  $\rho = \infty$  is shown in Fig. 4, while a transition region of finite extent is depicted in Figure 5.

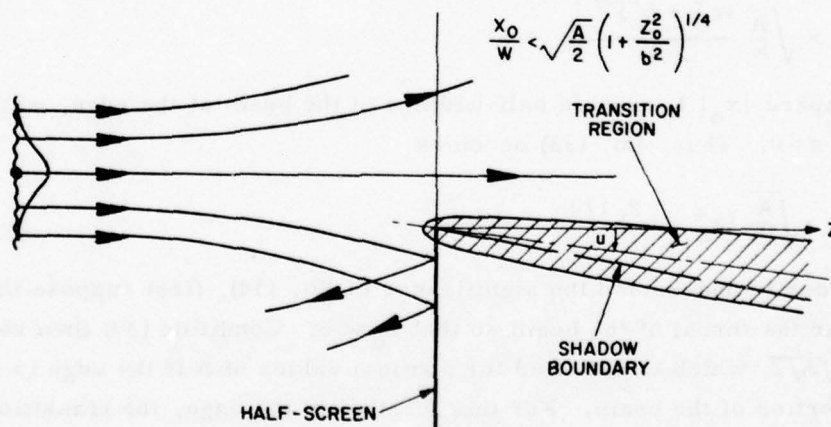


Fig. 4. Transition region about the incident-field shadow boundary for the case when the transition region extends to  $\rho \rightarrow \infty$ .

When the edge is in the paraxial region of the beam,  $|v| \ll 1$  and  $|\rho'|^2 \approx z_0^2 + b^2$  so that with the help of Eq. (18), condition (32) for the transition region to be finite extent becomes

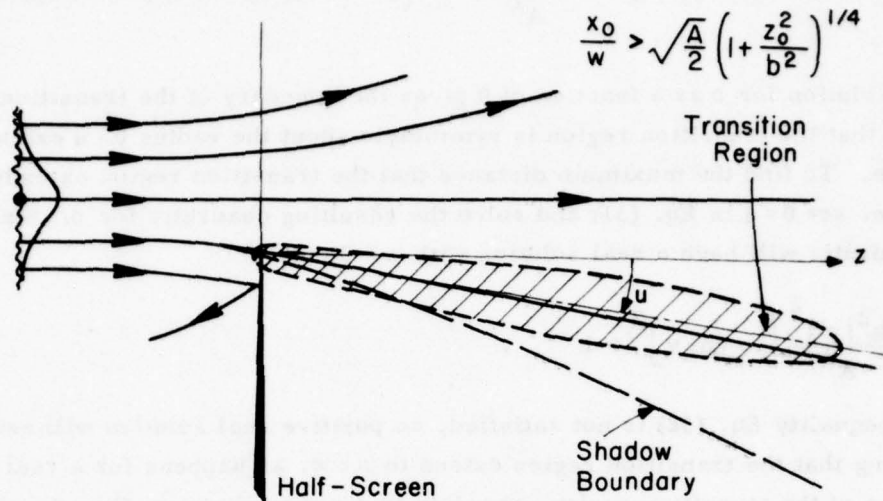


Fig. 5. Transition region about the incident-field shadow boundary for the case when the transition region is of finite extent along the radius  $\theta = u$ .

$$|x_0| > \sqrt{\frac{A}{k}} \frac{(z_0^2 + b^2)^{3/4}}{b} \quad (33)$$

Let us compare  $|x_0|$  to the  $1/e$  half-width  $w$  of the beam at the edge, which is given by (6) with  $z = 0$ . Thus, Eq. (33) becomes

$$\frac{|x_0|}{w} > \sqrt{\frac{A}{2b}} (z_0^2 + b^2)^{1/4} \quad (34)$$

In order to understand the significance of Eq. (34), first suppose the edge is located near the throat of the beam so that  $z_0 \ll b$ . Condition (34) then reduces to  $|x_0|/w > \sqrt{A/2}$ , which is satisfied for nominal values of  $A$  if the edge is outside of the main portion of the beam. For this location of the edge, the transition region has finite radial extent away from the edge, as depicted in Figure 5. However, for  $|x_0| \rightarrow 0$ , condition (34) is not satisfied for any value of  $A$  and the transition regions extends to  $\rho \rightarrow \infty$ , as in Figure 4. The discussion given above for  $z_0 \ll b$  also holds qualitatively for  $z_0 \sim b$ .

For  $z_0 \gg b$ , condition (34) becomes  $|x_0|/w > \sqrt{Az_0/2b}$  and is satisfied only for  $x_0$  much larger than the local beam width  $w$ . In this case the field incident on the edge  $E_1(0; \rho')$ , and hence the diffracted fields are much smaller than the field on the beam axis. When  $|x_0|/w < \sqrt{z_0 A/2b}$ , the transition regions extend to  $\rho \rightarrow \infty$ , as in

Figure 4. For  $z_0 \gg b$ , the shadow boundary  $\theta_{si}$  approaches  $\approx \tan^{-1}(x_0/z_0)$ . Thus the transition region becomes centered about the shadow boundary, which is the extension of the straight line from the beam center  $(-x_0, -z_0)$  passing through the edge. This behavior is the same as would be obtained for a line source, which is to be expected for  $z_0 \gg b$  since the edge is then located in the far field region of the source where the radiated fields are those of a line source multiplied by a slowly varying pattern function.

When the transition region does extend to  $\rho \rightarrow \infty$ , it occupies a region of finite angular width for  $\rho \rightarrow \infty$ . The width can be found from Eq. (31) by retaining only the highest powers in  $\rho$ . Using small argument expansions for the trigonometric and hyperbolic functions, it is found that the angular width  $\Delta\theta$  of the transition region is

$$\Delta\theta = 2\sqrt{\frac{A}{k|\rho|} - v^2} \quad (35)$$

Depending on the location of the edge, the transition region may cover the shadow boundary and the beam axis.

The foregoing discussion of the transition region about the shadow boundary of the incident field also holds for the reflected field transition region. The reflected field transition can be bound by inverting the incident field transition region through the plane  $z = 0$ .

Joint Services Technical Advisory Committee  
F44620-74-C-0056

A. C. Green, H. L. Bertoni and L. B. Felsen

U. S. Army Research Office  
DAHC04-75-G-0152

#### REFERENCES

1. G. A. Deschamps, "The Gaussian Beam as a Bundle of Complex Rays," *Electronics Letters*, 7 (1971).
2. S. Y. Shin and L. B. Felsen, "Gaussian Beam Modes by Multipoles with Complex Source Points," *JOSA*, 67 (1977).
3. G. Otis, "Application of the Boundary-Diffraction-Wave Theory to Gaussian Beams," *JOSA*, 64, pp. 1545-1550 (1974).
4. V. Shevernev, "Diffraction of a Plane Inhomogeneous Wave by a Semi-Plane," *Radiophysica*, 19, pp. 1854-1861 (1976) (in Russian).
5. H. L. Bertoni, A. C. Green and L. B. Felsen, "Shadowing of an Inhomogeneous Plane Wave by an Edge," *JOSA*.
6. J. J. Bowman, et al., "Electromagnetic and Acoustic Scattering by Simple Shapes," North-Holland Publishing Co., Amsterdam, pp. 323-327 (1969).
7. L. B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," Prentice-Hall, Inc., Englewood Cliffs, New Jersey, pp. 639-643 (1973).

# SCATTERING OF SURFACE WAVES BY A DIELECTRIC STEP DISCONTINUITY: OBLIQUE INCIDENCE CASE

J. P. Hsu, S. T. Peng and A. A. Oliner

A step discontinuity between two uniform dielectric waveguides occurs naturally in many problems of current interest. For example, a dielectric stripline for either millimeter wave or integrated optics applications may be viewed as consisting of two step discontinuities at the side walls, and a guided wave may be viewed as being reflected back and forth by the step discontinuities as it propagates along the stripline, as shown in Figure 1. Also, with the mathematical rigor attainable in such an electro-

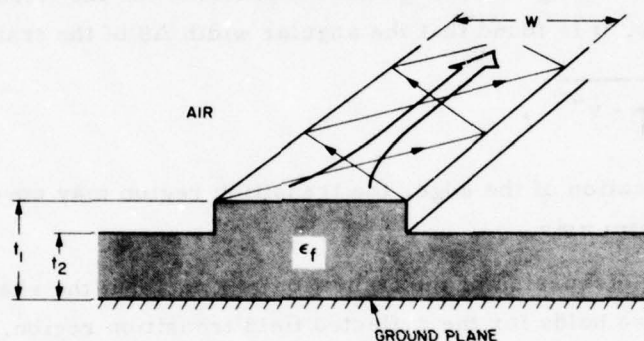


Fig. 1. Guidance along a dielectric ridge waveguide.

magnetic boundary value problem, many practical problems that are not amenable to a rigorous analysis can be approximated by a structure with one or more step discontinuities, known as the staircase approximation. Such an approach has been employed in a study of the radiation characteristics of high gain dielectric taper antennas of the type shown in Figure 2.

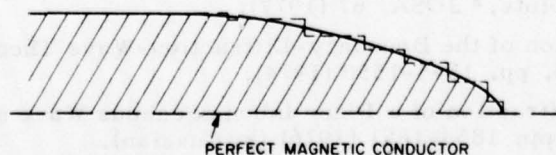


Fig. 2. Dielectric taper antenna -- symmetric bisection. The staircase approximation, involving a series of dielectric steps, is shown.

In the literature, the analysis of the step discontinuity has been restricted to the special case of normal incidence, where the TE and TM surface-wave modes exist

independently.<sup>1-3</sup> On the other hand, at an oblique incidence angle, both TE and TM modes must be simultaneously present in order to satisfy the required boundary conditions at a step discontinuity. In other words, for either TE or TM mode incidence, both modes are excited at the step discontinuity. Such mode coupling may result in interesting physical phenomena, such as a novel resonance effect and leakage of energy from a dielectric waveguide.<sup>4</sup>

The difficulty encountered in solving the class of dielectric waveguides which are open structures is in the handling of the continuous spectrum that accounts for the radiation and storage of electromagnetic energy in the presence of discontinuities. In the special case of normal incidence on the step junction, a rigorous treatment of the continuous spectrum requires the solution to an integral equation which is not amenable to an exact analysis.<sup>1</sup> In the general case of oblique incidence, it is expected that an exact analysis will result in a system of integral equations for which the construction of an accurate solution is even harder. Instead of approximating the governing equations, we enclose the dielectric stripline in a metallic waveguide so that the mode functions for the representation of the electromagnetic fields become discrete. In such a closed system, the mode amplitudes are determined from a system of linear equations from which the numerical results can be obtained with known accuracy. Therefore, the novel wave phenomena previously predicted can now be analyzed accurately.

A step discontinuity between two dielectric waveguides is shown in Figure 3. The

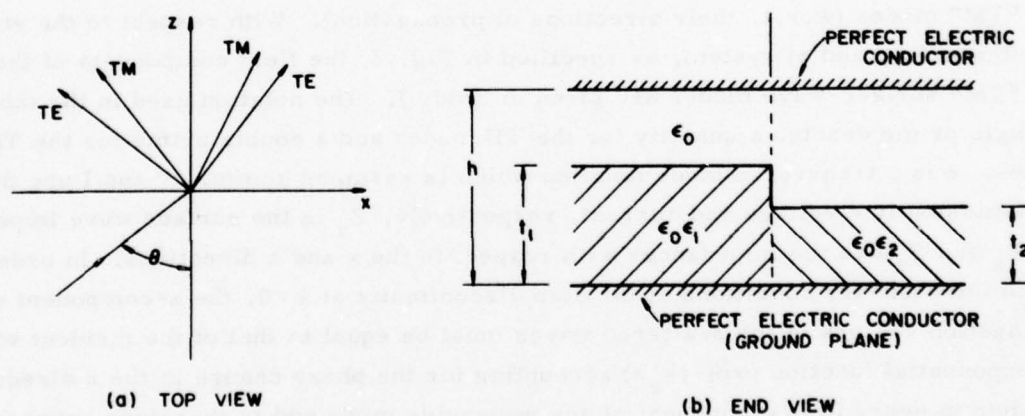


Fig. 3. Scattering of a surface wave by a dielectric step discontinuity. The top view in (a) shows that incidence at an oblique angle produces coupling between TE and TM modes. The structure shown in end view in (b) includes the fictitious top perfectly-conducting plate introduced to aid in the calculations by discretizing the continuous spectrum.

discontinuity may be due to a change in the physical dimensions ( $t_1 \neq t_2$ ) or the material property ( $\epsilon_1 \neq \epsilon_2$ ) or both. Here, we assume that the structure is uniform and infinite in extent along the  $z$ -direction. For most millimeter wave applications, the dielectric waveguides are expected to be on a ground plane, as shown in Figure 3. Such a structure may also be considered as the short-circuit bisection of a larger structure that is symmetric in  $y$ . The upper metallic plate at the height  $y = h$  is introduced artificially in order to discretize the mode functions so that the ensuing analysis can be carried out exactly. The mode functions in a partially filled parallel plate waveguide are well-known; they consist of surface-wave modes that are evanescent in the air region and non-surface-wave modes that vary sinusoidally in the transverse ( $y$ ) direction. Since the incident wave is expected to be a surface-wave mode which is evanescent in the air region, the effect of the upper metallic plate at a sufficiently large height should be exponentially small. For the original open structure (without the upper metallic plate), the energy radiated away from the discontinuity can now be accounted for by the propagating non-surface-wave modes of the closed waveguide and the energy stored in the vicinity of the discontinuity can be accounted for by those modes that are below cutoff in the  $x$ -direction in the metallic waveguide. Therefore, in this approach, it is necessary to evaluate the amplitudes of the propagating as well as the evanescent modes.

It is shown in Fig. 3(a) that a surface-wave mode is incident obliquely at an angle  $\theta$  onto a step discontinuity between two uniform dielectric waveguides. In contrast to the special case of normal incidence ( $\theta = 0$ ), the general case of oblique incidence is a vector boundary value problem that requires the simultaneous presence of both "TE" and "TM" modes (w.r.t. their directions of propagation). With respect to the structure coordinate ( $x, y$  and  $z$ ) system, as specified in Fig. 3, the field components of the "TE" and "TM" surface-wave modes are given in Table I. The notation used in the table is: A single prime denotes a quantity for the TE modes and a double prime for the TM modes.  $\phi$  is a transverse mode function which is assumed known.  $V$  and  $I$  are the transmission line voltage and current, respectively,  $Z_s$  is the surface wave impedance, and  $Z_x$  and  $Z_z$  are the impedances with respect to the  $x$  and  $z$  directions. In order to match the boundary conditions at the step discontinuity at  $x = 0$ , the  $z$ -component of the propagation vectors of the scattered waves must be equal to that of the incident wave; the exponential function  $\exp(-jk_z z)$  accounting for the phase change in the  $z$  direction is common to every field component of any waveguide mode and is therefore suppressed in the table for simplicity. The continuity of the four tangential field components ( $E_y$ ,  $E_z$ ,  $H_y$  and  $H_z$ ) results in the following four systems of linear equations that can be written in the matrix form as:

TABLE I. Field components and their representations.

TE mode	TM mode
$E'_x = Z'_z I'(x) \phi'(y)$	$E''_x = -I''(x) \left[ \frac{1}{j\omega\epsilon_0\epsilon(y)} \frac{d}{dy} \phi''(y) \right]$
$E'_y = 0$	$E''_y = V''(x) \phi''(y) \frac{1}{\epsilon(y)}$
$E'_z = -V'(x) \phi'(y)$	$E''_z = Y''_z V''(x) \left[ \frac{1}{j\omega\epsilon_0\epsilon(y)} \frac{d}{dy} \phi''(y) \right]$
$H'_x = V'(x) \left[ \frac{1}{j\omega\mu} \frac{d}{dy} \phi'(y) \right]$	$H''_x = Y''_z V''(x) \phi''(y)$
$H'_y = I'(x) \phi'(y)$	$H''_y = 0$
$H'_z = Z'_z I'(x) \left[ \frac{1}{j\omega\mu} \frac{d}{dy} \phi'(y) \right]$	$H''_z = I''(x) \phi''(y)$
$\frac{d}{dx} V'(x) = -jk'_x Z'_x I'(x)$	$\frac{d}{dx} V''(x) = -jk''_x Z''_x I''(x)$
$\frac{d}{dx} I'(x) = -jk'_x Y'_x V'(x)$	$\frac{d}{dx} I''(x) = -jk''_x Y''_x V''(x)$
$Z'_x = 1/Y'_x = Z_s \sin \theta'$	$Y''_x = 1/Z''_x = Y_s'' \sin \theta''$
$Z'_z = 1/Y'_z = Z_s \cos \theta'$	$Y''_z = 1/Z''_z = Y_s'' \cos \theta''$
$Z'_s = \omega\mu_0/\kappa'$	$Y''_s = \omega\epsilon_0/\kappa''$

$$\underline{I}'(0) = P' \underline{\bar{I}}'(0) \quad (1)$$

$$\underline{V}''(0) = P'' \underline{\bar{V}}''(0) \quad (2)$$

$$R' \underline{I}'(0) + \underline{I}''(0) = S' \underline{\bar{I}}'(0) + Q'' \underline{\bar{I}}''(0) \quad (3)$$

$$\underline{V}'(0) + R'' \underline{V}''(0) = Q' \underline{\bar{V}}'(0) + S'' \underline{\bar{V}}''(0) \quad (4)$$

where  $\underline{I}'(0)$  and  $\underline{V}'(0)$  are infinite column vectors consisting of the TE modal current and voltage of the waveguide on the left and evaluated at  $x=0$ , and  $\underline{I}''(0)$  and  $\underline{V}''(0)$  are those for the TM modes. The quantities with a superbar are those for the waveguide on the right. The coefficient matrices  $P, Q, R$  and  $S$  with either a single prime or a double prime are related to the scalar products of the known mode functions. Evidently, Eqs. (3) and (4) include the effect of coupling between TE and TM modes. The analysis above is exact for the structure under consideration.

We intend to utilize the analysis discussed above to obtain numerical values for the scattering characteristics of a variety of different dielectric step discontinuities, particularly those which arise naturally as the sides of waveguiding structures. However, we have already obtained approximate numerical values for structures corresponding to some waveguides used in integrated optics and millimeter waves, under the condition that the continuous spectrum is neglected. Thus, only the surface wave modes are included in the solution, and the problem simplifies considerably. Those solutions will not yield information on the radiation pattern, but they will tell us what major effects are produced by the TE-TM mode coupling introduced by the oblique angle incidence. Later, using the complete solution above, we can determine the added influence of the continuous spectrum.

One of the step junctions analyzed is the side of the inverted strip waveguide for millimeter waves, a structure proposed by Itoh. As an example of those numerical calculations, which were obtained for a variety of parameters values, we present in Fig. 4 the scattering characteristics obtained for TE mode incidence, since they illus-

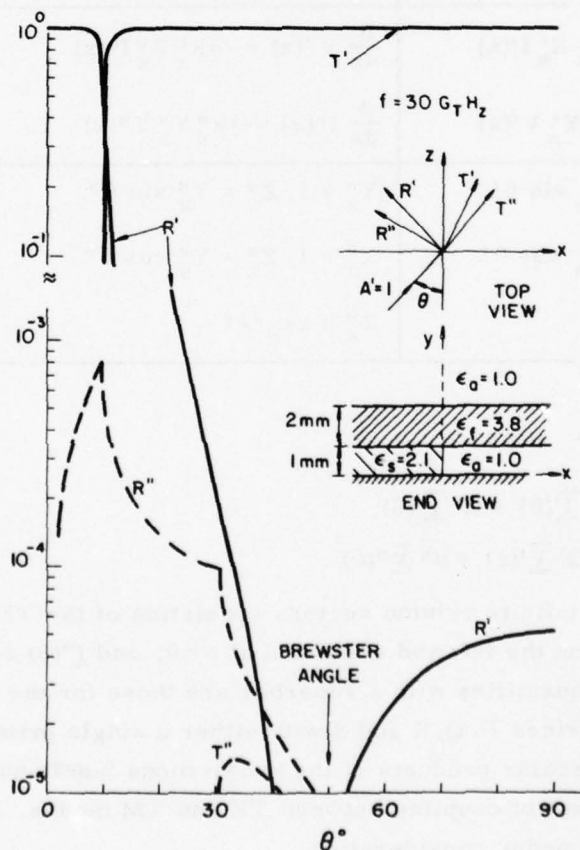


Fig. 4. Relative amplitudes of the reflected and transmitted surface wave modes when a TE surface wave mode is incident at an oblique angle  $\theta$  on the dielectric step junction shown (corresponding to a side of the inverted strip waveguide for mm waves). Note the presence of a null corresponding to a Brewster angle.

trate an interesting physical effect. The physical dimensions of the structure are shown in the insert, and are chosen such that each of the two uniform waveguiding regions supports only the fundamental TE and TM modes at the frequency of 30 GHz. For the fundamental TE surface wave incident from the right of the discontinuity, the relative amplitudes of the scattered fundamental modes are shown as a function of the incident angle  $\theta$ . From Fig. 4, we observe that there exists a null (near  $\theta = 50^\circ$ ) in the specular reflection coefficient, as the incident angle is scanned. No such null reflection is observed for the other polarization. Such a null reflection for "TE" surface wave incidence may be interpreted as the analogue of the Brewster angle effect which is known to exist in the case of TM plane wave scattering by an interface between two dielectric media. Possible practical implications of the Brewster angle effect on waveguiding characteristics are still being investigated.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

J. P. Hsu, S. T. Peng and A. A. Oliner

#### REFERENCES

1. S. T. Peng and E. W. Hu, "Analysis of Dielectric Waveguide Discontinuities," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, pp. 61-65 (November 1975).
2. C. M. Angulo, "Diffraction of Surface Wave by a Semi-Infinite Dielectric Slab," Doctoral Dissertation, Polytechnic Inst. of Brooklyn (1955).
3. S. F. Mahmoud and J. C. Beal, "Scattering of Surface Waves at a Dielectric Discontinuity on a Planar Waveguide," IEEE Trans., Vol. MTT-23, pp. 193-198 (1975).
4. S. T. Peng and A. A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics," International Conf. on Integrated Optics and Optical Fiber Communication, Tokyo, Japan (July 1977).

## NEW PROPAGATION EFFECTS FOR THE INVERTED STRIP DIELECTRIC WAVEGUIDE FOR MILLIMETER WAVES

A. A. Oliner, S. T. Peng and J. P. Hsu

Associated with the resurgence in interest during the past few years in millimeter waves, several novel waveguiding structures have been proposed which are suitable for millimeter wave integrated circuits but which avoid the high loss associated with the metallic strip of microstrip line at these high frequencies. One of the most promising of these is the inverted strip dielectric waveguide,<sup>1,2</sup> the cross section of which is shown in Figure 1. This waveguide, which was proposed by Itoh, has the added virtue that the loss due to the metallic ground plane is also minimized, since the dielectric constant values are so chosen that most of the energy resides in the upper dielectric sheet. The dielectric strip serves to confine the guided wave laterally, but the field in the strip region is evanescent vertically, with the result that the field becomes small when it finally reaches the metallic ground plane, and the ground plane current remains low as a consequence.

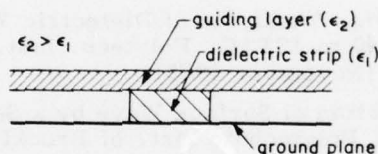


Fig. 1 Cross section of the inverted strip dielectric waveguide for millimeter waves.

The published theoretical analysis<sup>1</sup> for the propagation characteristics of this waveguide type is approximate, but it serves well enough to produce reasonable agreement with available experimental results.<sup>1</sup> However, when a more careful examination is made of the comparison between theory and measurements, it is seen that the agreement is poor in some regions and that, in fact, some puzzling discrepancies appear. We have conducted an improved analysis of the behavior of the structure, and we have shown that this improved theory not only produces better agreement with measurement but also reveals a number of very interesting and fundamental propagation effects which the earlier approximate analysis misses entirely.

The approximate theoretical analysis<sup>1</sup> mentioned above utilizes the so-called "equivalent dielectric constant" (edc) method, which is essentially a simplified transverse resonance procedure that assumes the presence of only one mode, the lowest TM mode, in each of the constituent transverse regions comprising the guide cross section. Furthermore, in the transverse resonance taken under these conditions in the horizontal direction, the geometrical discontinuities present at the strip sides are completely

neglected since each constituent region is viewed as possessing an "equivalent dielectric constant."

The first thing that a more accurate analysis reveals is that in each constituent region in the guide's cross section at least two propagating modes, one TM and the other TE, must be present simultaneously under almost all conditions. In fact, if the dielectric sheet or dielectric strip thickness were increased, then the number of modes present simultaneously would increase to four or to six, etc. When we examine the three structures actually analyzed and measured by Itoh,<sup>1</sup> we find that four simultaneous modes are present in two of them, and that six simultaneous modes are present in part of the third guiding structure.

The second key point that a more accurate analysis indicates is that the TE and TM modes which are present necessarily couple to each other at the geometrical discontinuities corresponding to the sides of the strip. This coupling requirement is readily proven by an examination of the field components that must be present at these strip sides.

The simultaneous presence of these other modes in the constituent regions, and their coupling at the strip sides, produce a number of interesting and sometimes important physical effects now observed earlier. Among these effects are the following:

- (a) Under appropriate conditions, leakage from the strip sides can occur, changing the dominant bound mode into a leaky mode.
- (b) Under appropriate conditions, for certain frequencies or strip widths, resonance effects can occur.
- (c) The transverse field behavior, both vertically and horizontally, can differ significantly from what one expects on the basis of the dominant mode alone.
- (d) If discrete modes higher than the lowest TE and TM ones are present on the line, they will almost always be leaky.

These previously unexpected propagation effects could produce unwanted performance difficulties under some circumstances, or else advantage could be taken of them if it is known in advance what the effects are and when they occur. Furthermore, by a proper choice of dimensions or dielectric constants these effects can either be made to appear or be eliminated.

We present in Figure 2 calculations for the basic surface wave characteristics of the constituent regions comprising the inverted strip guide cross section. The calculations are in the form of the "effective dielectric constant,"  $\epsilon_{\text{eff}}$ , which is proportional to the propagation constant, as a function of the dimensions for the specific parameters chosen by Itoh<sup>1</sup> for his three waveguides. (Actually, the plots are given as  $n_{\text{eff}} = (\epsilon_{\text{eff}})^{1/2}$  vs.  $h$ , which is the height of the strip, when  $d$ , the height of the upper layer,  $f$ , the frequency, and the dielectric constant values in each region are specified.)

The two curves labeled  $\epsilon_1 = 1.0$  correspond to the constituent portions of the guide cross section which are outside of the strip portion, whereas the two curves labeled  $\epsilon_1 = 2.1$  hold for the strip region itself. The three cases examined by Itoh correspond to  $h = 0.794$  mm, 1.588 mm and 3.175 mm. Both inside and outside of the strip region, at least 4 modes are present simultaneously, 2 TE modes and 2 TM modes. In his analysis, Itoh assumed that only the lowest TM mode was present.

Several other interesting features also follow from Figure 2. Note that the lowest TM mode, which is the mode incident and the one which contains most of the power, is no longer the "dominant" mode in the sense that it is the slowest mode, as it is for a single dielectric layer on a ground plane. As  $h$  increases, a cross-over occurs, and the TE mode becomes the "dominant" one. It is the occurrence of this interesting cross-over that makes possible the leakage and the resonance effects.

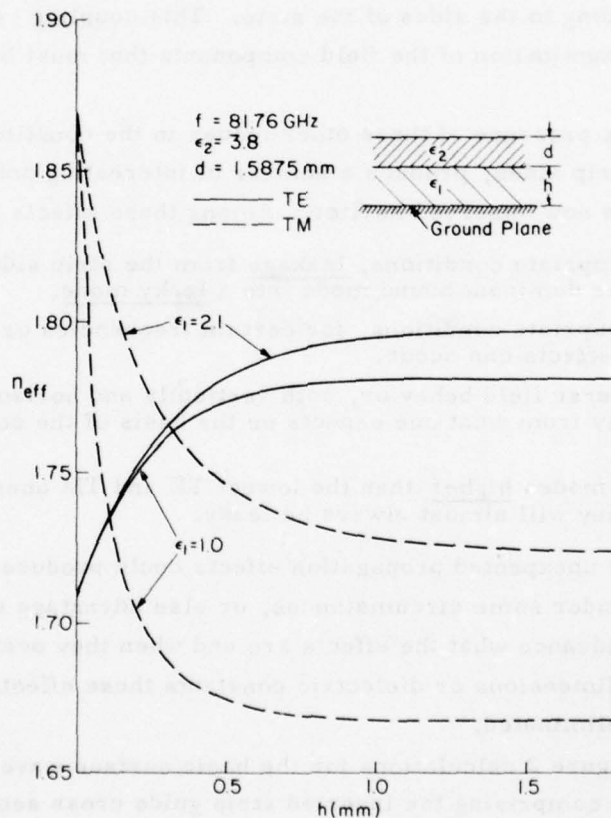


Fig. 2. Basic surface wave characteristics for the constituent regions comprising the guide cross section. "Effective dielectric constant" values as a function of strip height for all of the discrete modes existing at that frequency (81.7 GHz).

The curves in Figure 2 are used in a simple way to determine when these leakage and resonance effects will appear. The procedure follows the one developed for a similar purpose in the context of integrated optics.<sup>3</sup> As mentioned earlier, the leakage and resonance effects arise because of the TE-TM coupling which occurs at the strip sides. The TM wave inside the strip region, which is incident at an angle on the strip side, excites a reflected propagating TM wave inside, an evanescent TM wave outside, and (due to the coupling) TE waves both inside and outside the strip region. If the TE wave excited outside is propagating rather than evanescent, it carries away a small amount of energy at each reflection, resulting in a leaky mode rather than a purely bound mode. If the TE wave inside is excited at an appropriate angle, then resonance effects can occur at certain strip widths. (Similar leakage and resonance behaviors have been reported for strip waveguides for integrated optics,<sup>3</sup> but certain differences are present in this millimeter wave structure.)

One or the other of these special effects can be present (or absent) depending on the precise dimensions of the structure. Table I summarizes the resonance and leakage behavior for three sets of dimensions for the inverted strip waveguide, using the dielectric constant values employed by Itoh.

TABLE I

Basic mode: lowest TM

$$\epsilon_1 = 2.1, \epsilon_2 = 3.8$$

Case	$d/\lambda$	$h/\lambda$	Resonance	Leakage
1	0.10	0.20	no	yes, if $W/\lambda < 0.3$
2	0.20	0.15	yes	yes, if $W/\lambda < 1.5$
3	0.10	0.10	no	no

For case 1, we see that leakage can occur without resonance effects. However, the leakage may be eliminated if the strip width is increased beyond about  $0.3\lambda$ , which is not difficult to accomplish. On the other hand, in case 2 we observe both resonance and leakage, where the leakage would be present over a rather wide range of values of  $W/\lambda$ . In case 3, neither effect is present, showing that it is certainly possible to design the waveguide without encountering these problems.

## REFERENCES

1. T. Itoh, "Inverted Strip Dielectric Waveguide for Millimeter-Wave Integrated Circuits," IEEE Trans. Microwave Theory Tech., Vol. MTT-24, pp. 821-827 (November 1976).
2. R. Rudokas and T. Itoh, "Passive Millimeter-Wave IC Components Made of Inverted Strip Dielectric Waveguides," IEEE Trans. Microwave Theory Tech., Vol. MTT-24, pp. 978-981 (December 1976).
3. A. A. Oliner and S. T. Peng, "Leaky Modes on Optical Strip Waveguides," International Symposium on Optical Communication, 19th URSI General Assembly, Helsinki, Finland (August 1-8, 1978).
4. S. T. Peng and A. A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics," Trans. Inst. of Electronics and Communication Engineers of Japan, Special Issue on Integrated Optics and Optical Fiber Communications, Vol. E61, pp. 151-154 (March 1978).

## INTERACTIVE MECHANISMS AND EFFECTS OF LOW-LEVEL MILLIMETER WAVES ON LIVING SYSTEMS

S. Motzkin, S. W. Rosenthal, L. Birenbaum, R. Melnick, C. Rubenstein, R. Remilly and S. Davidow

Preliminary results of an experimental study of the effects of single CW mm - wave exposure on the physiological and metabolic processes of prokaryotic and eukaryotic cells are reported.

A millimeter wave exposure system has been developed to be used over a frequency range of 26.5-75GHz and which can further be extended to 110 GHz. The system consists of three separate waveguide sizes, RG-96/U or WR-28 (Figs. 1, 2) (26.5-40 GHz, 11.3 - 7.5 mm), RG - 97/U or WR - 22 (Figs. 3, 4) (33-50 GHz, 9.1-6.0mm), and RG-98/U or WR - 15, (50-75 GHz, 6.0 - 4 mm). The above systems terminate in a horn, the aperture of which is placed about 2 mm above the experimental material to be exposed. Biological samples are either condensed on a filter paper surface or in solution in a water-jacketed, temperature controlled plexiglass cell (Figures 5, 6). Standard millimeter waveguide components are used for dosimetry and reproducibility in this system and thus insure a known exposure of the cellular material (Figures 7, 8).

Cells and subcellular components which are currently under investigation include E. Coli (strains W3110 with Col E<sub>1</sub> factor and W1485, rat liver mitochondria and red blood cells. Earlier work, carried out by scientists in the Soviet Union<sup>4-6, 9, 10, 12-16</sup> near 50 GHz at low power levels revealed specific frequency dependent effects which are cyclical and suggest resonance phenomena.

#### A. E. Coli Experiments

Colicin, an antibiotic, is synthesized in bacterial strains which have a plasmid that contains colicinogenic factor. Although present, Col E<sub>1</sub> cells normally do not actively produce large quantities of colicin. However, if these cells are incubated above 43°C, treated with mitomycin C, or exposed to ultraviolet light, then active colicin synthesis and replication of DNA molecules which code for Col E<sub>1</sub> occurs. Russian scientists have reported that millimeter waves can stimulate increased colicin production in a frequency dependent manner, which is almost insensitive to power densities over 2 orders of magnitude.<sup>15</sup> Attempts to reproduce these experiments are in progress.

Col E<sub>1</sub> cells in log and in stationary phase are to be irradiated at intervals in the 51.3-52.2 GHz frequency range (5.85 to 5.75 mm) for 30 to 60 minutes, at incident power densities of approximately .001 mW/cm<sup>2</sup> at 25°, 37°, and 43°C.

Col E<sub>1</sub> cells are irradiated in a water-jacketed, temperature controlled cell (Figure 5). Exposed Col E<sub>1</sub> cells are grown, concentrated and thinly spread over a nutrient agar plate to achieve a homogeneous lawn. These cells are chloroformed and Col sensitive bacteria are then inoculated onto the plate (Figure 9). Clear areas or plaques (Fig. 10), which develop during the subsequent growth period are scored. These areas are indicative of the destruction of Col sensitive bacteria. Results in irradiated specimens are compared with controls of Col E<sub>1</sub> cells grown alone, and of uninduced Col E<sub>1</sub> and sensitive bacteria grown together (Figure 11). How the amount of colicin is related to the plaque size has yet to be determined.

Preliminary observations indicate that colicin production is enhanced at specific frequencies at 25° and 37°C. Earlier investigators have reported that temperature above 43° induces colicin production. Below this threshold, increases induce colicin production only at specific frequencies. Even in unirradiated controls some colicin production is evident. In experiments reported here, colicin production is doubled in 60 minutes at 37°C and a power density of .5mW/cm<sup>2</sup> following irradiation.

#### B. Experiments with Mitochondria

Although the mechanism by which weak electromagnetic fields act on membranes is unknown, several suggestions have been advanced including: modification of charged binding sites by conformational changes or displacement of surface bound ions,<sup>3</sup> transformations of the hydrated phase of globular proteins<sup>11</sup> (which are strongly dependent on divalent cations and which could in turn propagate or amplify local electrical events), changes in membrane potential,<sup>2</sup> and stimulation of coherent molecular oscillations.<sup>7,8</sup> Mitochondrial membranes are being utilized in these experiments because they are well characterized. Exposing them to millimeter waves will permit us to define possible alterations in their macromolecular interactions and to determine the significance of these variations on the functional capacity of living cells.

These organelles, extracted from rat liver cells, and centrifuged to a paste-like consistency with a concentration of 70 mg/ml protein, are placed in a water-jacketed accuvette (Fig. 6) for exposure. Irradiations carried out at 35 GHz (8.6 mm), in preliminary studies, included relatively high power densities. Studies now in progress will include 53.7-50.5 GHz (5.59-5.94 mm) as well as varying power densities from 0.01-1.0 mW/cm<sup>2</sup> for 12-15 minutes. Metabolic membrane activity is evaluated by measuring total ATP synthesis as a function of oxidative phosphorylation. A technique has been devised which will make it possible to irradiate mitochondria in a physiologically active state without the interference of electrodes imposed in the

electromagnetic field. During the latter procedure, the concentration of experimental material is reduced to 1.5 mg/ml to increase the duration of experimental time available before essential reactants (e.g. oxygen) are used up. ATP synthesized is trapped as glucose-6-phosphate using glucose and hexokinase to prevent ATP hydrolysis. Subsequently, glucose-6-phosphate concentration is determined by evaluating NADP reduction with glucose-6-phosphate dehydrogenase at 340 nm. Membrane integrity is evaluated as a function of oxidative phosphorylation which is a complex process involving several components in the electron transport system and the ATPase complex. Alterations in the membrane which affect one component will affect the entire process.

To date, drastic changes in respiratory control have not been observed by us except with 35 GHz at very high power densities where irreversible membrane injury has certainly occurred. Preliminary results show no effects at 35GHz, at power densities of 10 mW/cm<sup>2</sup> or less. In experiments where effects have not been observed, radiation injury incurred could possibly have been repaired. Therefore, irradiations and analyses will be carried out on actively phosphorylating mitochondria so that evaluations can be accomplished prior to repair.

#### C. Experiments with Red Blood Cells

To further elucidate the effects of irradiation on cellular membranes, preliminary studies of red blood cells have been undertaken. It has previously been reported that 5-10 mW/cm<sup>2</sup> change the permeability of blood cell membranes to hemoglobin.<sup>1</sup> One ml of packed red blood cells in a water-jacketed, temperature-controlled cell, is being irradiated at 51.7, GHz (5.8 mm) for 15, 30, 60, 120 minutes at 7 mW/cm<sup>2</sup> and will be irradiated at 1 and 0.1 mW/cm<sup>2</sup> power density levels. Changes in the permeability and porosity of the membranes are being investigated by examining for potassium ion and hemoglobin leakage. Exposed cells are centrifuged and the supernatant fluid examined spectrophotometrically at 576 nm for hemoglobin absorption and with a potassium electrode to determine the concentrations of these components liberated from exposed cells. Results will be compared with those of controls handled in the same way but not irradiated.

Public Health Service  
FDA-1862-77-C

S. Motzkin

Office of Naval Research  
N00014-77-C-0413

#### REFERENCES

1. S. Baranski, S. Szmegielski and J. Moneta, Proc. Symp. Biological Effects and Health Hazards of Microwave Radiation, Polish Medical Publishers (1974).
2. F.S. Barnes and J.H. Chia-lun, "A Model for Some Non-Thermal Effects of Radio and Microwave Fields on Biological Membranes,"

3. S.M. Bawin, L.K. Kaczmarek and W.R. Adey, "Effects of Modulated VHF Fields on the Central Nervous System," *Annals of the New York Acad. of Sci.*, Vol. 247, pp. 74-81 (1975).
4. E.B. Bazanova, A.K. Bruykhova, R.L. Vilenskaya, E.A. Gelvich, M.B. Golant, N.S. Landau, M.V. Melnikova, N.P. Mikaelyan, G.M. Okhokhonia, L.A. Sevast'yanova, A.Z. Smolyanskaya and N.A. Sycheva, "Certain Methodological Problems and Results of Experimental Investigation of the Effects of Microwaves on Micro-organism and Animals," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
5. N.D. Devyatkov, "Influence of Millimeter-band Electromagnetic Radiation on Biological Objects," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
6. V.I. Gaiduk, Y.I. Khurgin and V.A. Kudryashova, "Outlook for Study of the Mechanisms of the Nonthermal Effects of Millimeter and Submillimeter-band Electromagnetic Radiation on Biologically Active Compounds," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
7. W. Grundler, F. Keilmann and H. Fröhlich, "Resonant Growth Rate Response of Yeast Cells Irradiated by Weak Microwaves," *Phys. Lett.*, Vol. 62A, No. 6, pp. 463-466 (1977).
8. W. Grundler and F. Keilmann, "Nonthermal Effects of Millimeter Microwaves on Yeast Growth," *Z. Naturforsch.*, Vol. 33C, pp. 15-28 (1978).
9. R.I. Kiselev and N.P. Zalyubovskaya, "Effects of Millimeter-band Electromagnetic Waves in the Cell and Certain Structural Elements of the Cell," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
10. V.F. Kondrat'eva, E.N. Christyakova, I.F. Shmakova, N.B. Ivanova, A.A. Treskurov, "Effects of Millimeter-band Radio Waves on Certain Properties of Bacteria," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
11. J.G. Llaurodo, A. Sances, Jr., J.H. Battocletti, "Biological and Clinical Effects of Low Frequency Magnetic and Electric Fields," *C.C. Thomas*, Ch. 13, pp. 172-186.
12. S.E. Manilov, E.N. Christyakova, V.F. Kondrat'eva and M.A. Strelkova, "Effects of Millimeter-band Electromagnetic Waves on Certain Aspects of Protein Metabolism Bacteria," *Scientific Session of the Division of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
13. *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973); *USP Fiz. Nauk.*, Vol. 110 (July 1973), pp. 452-469; also, *Sov. Phys. Usp.*, Vol. 16, No. 4, pp. 568-579 (1974).
14. L.A. Sevast'yanova and R.L. Vilenskaya, "A Study of the Effects of Millimeter-band Microwaves on the Bone Marrow of Mice," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
15. A.Z. Smolyanskaya and R.L. Vilenskaya, "Effects of the Millimeter-band Electromagnetic Radiation on the Functional Activity of Certain Genetic Elements of Bacterial Cells," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).
16. N.P. Zalyubovskaya, "Reactions of Living Organisms Exposure to Millimeter-band Electromagnetic Waves," *Scientific Session of the Div. of Gen. Phys. and Astronomy, USSR Acad. Sci.* (1973).

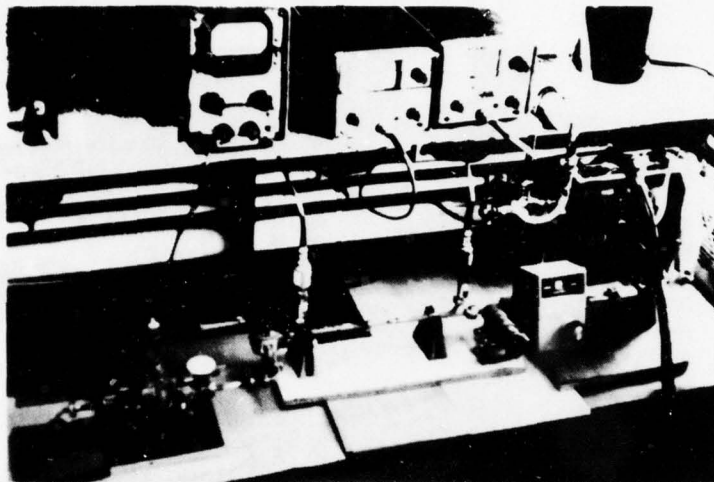


Fig. 1. Waveguide assembly and associated meters for a 35 GHz system.

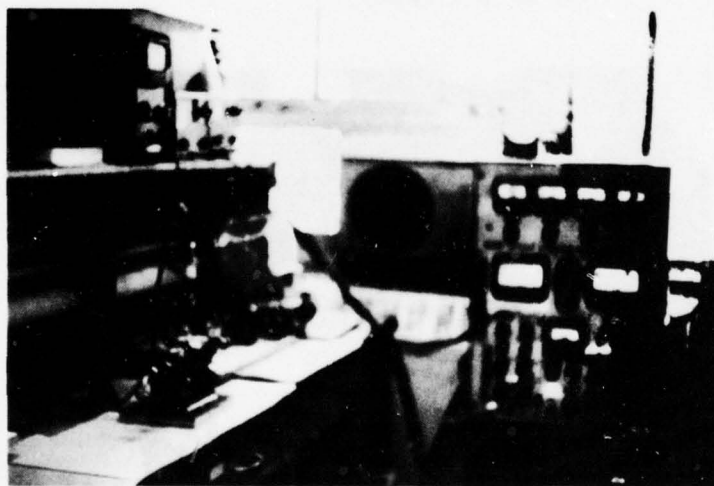


Fig. 2. Power supply for a 35 GHz klystron system.

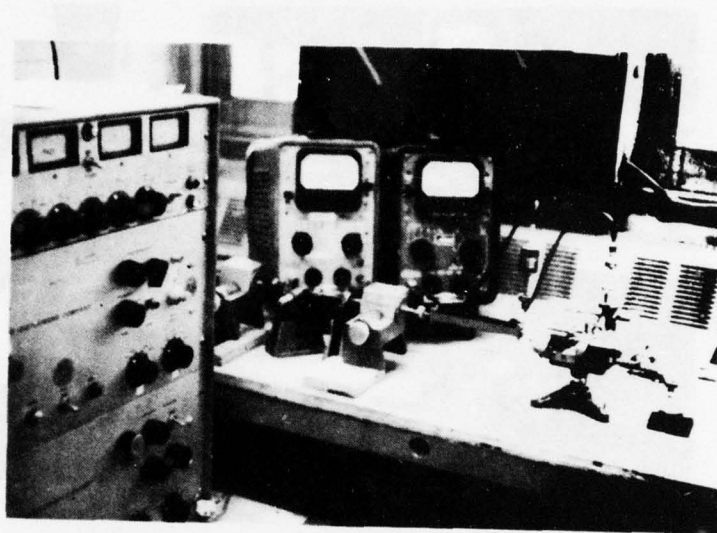


Fig. 3. Waveguide assembly and associated meters for BWO system.

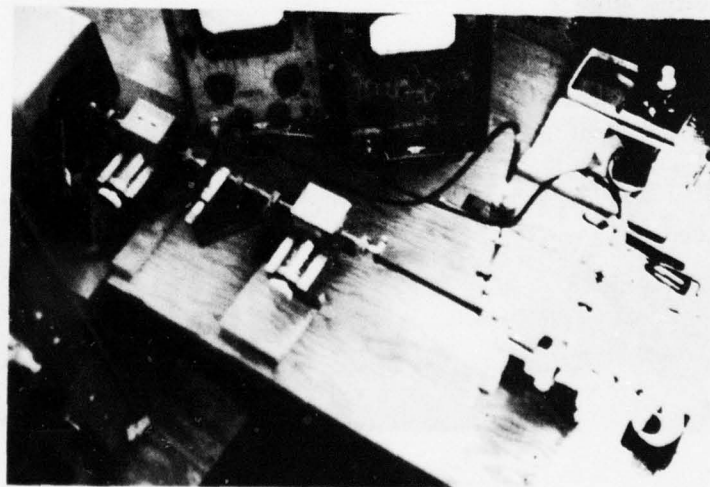
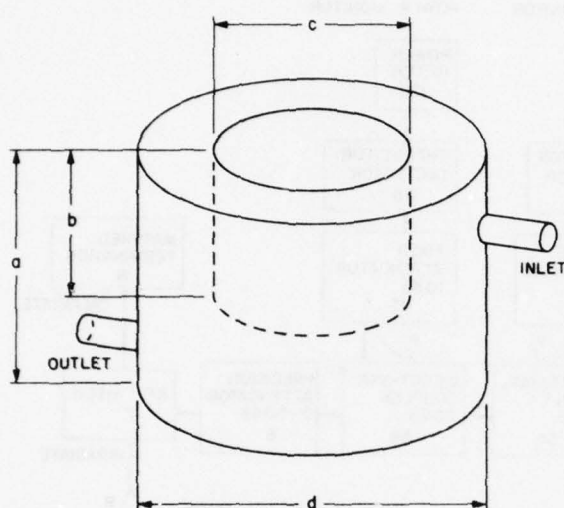


Fig. 4. Waveguide assembly and associated meters for BWO system.



Chamber I (inches)	Chamber II (inches)	Dimensions
2.50	4.50	a
1.75	2.625	b
0.75	2.50	c
4.25	5.125	d

NOTES:

1. The above is a prototype of plexi-glass chambers used in irradiations. Fabrication includes one with small Chamber I for red blood cells and a larger Chamber II to accommodate petri dish with E Coli.
2. Measurements of various dimensions for the two chambers are noted in inches.

Fig. 5. Water-jacketed sample chamber.

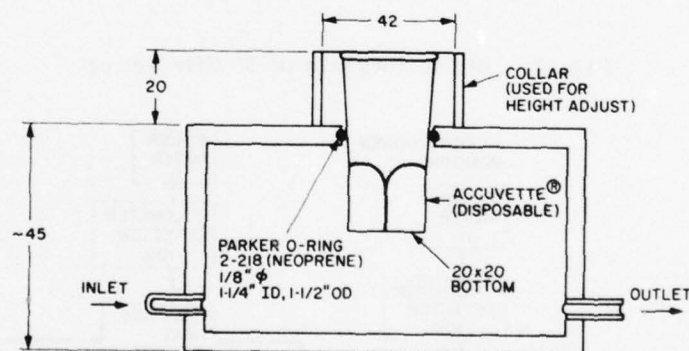


Fig. 6. Temperature controlled water jacket with accuvette.®

## ELECTROMAGNETICS

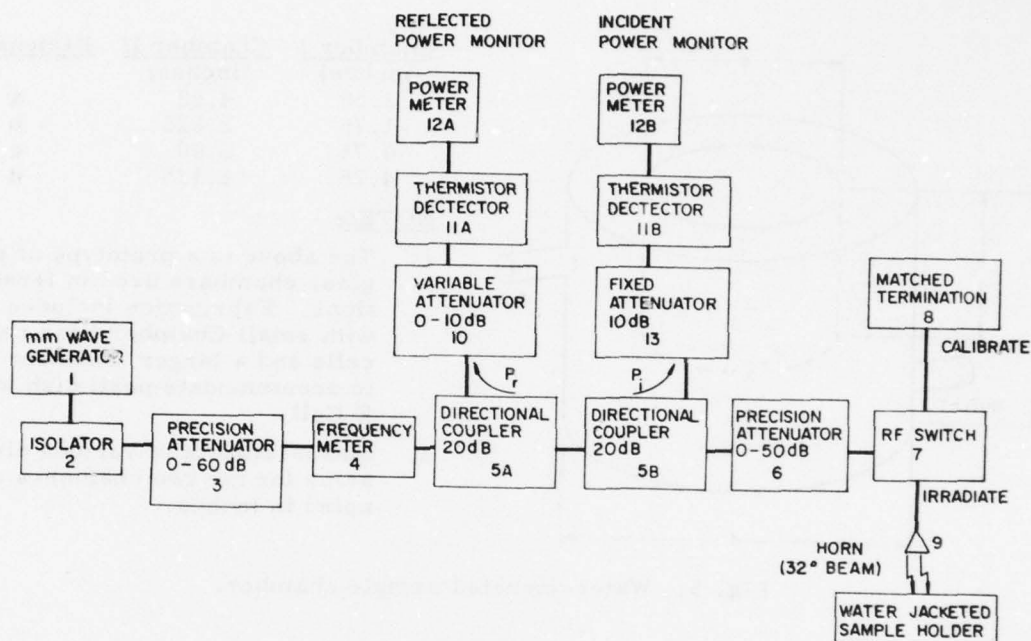


Fig. 7. Block diagram of 35 GHz set up.

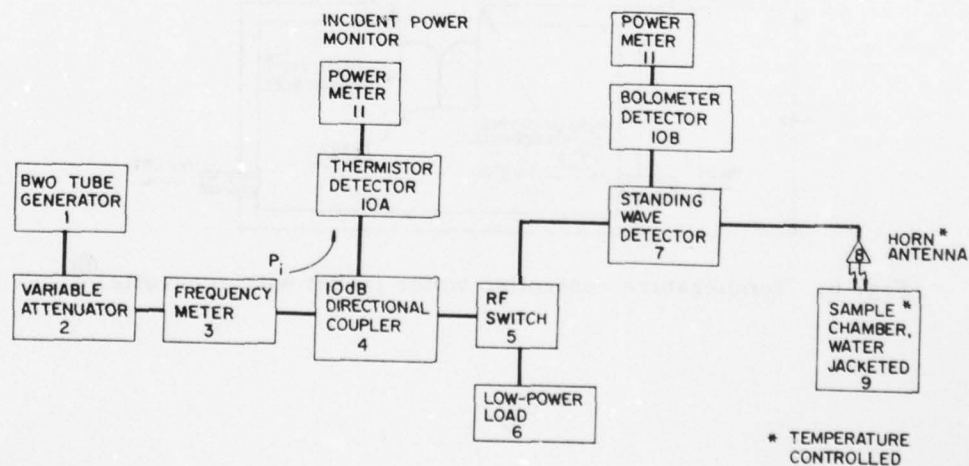


Fig. 8. Block diagram of BWO set up.

AD-A063 181

POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 9/3  
PROGRESS REPORT NUMBER 43 TO THE JOINT SERVICES TECHNICAL ADVIS--ETC(U)  
NOV 78 A A OLINER

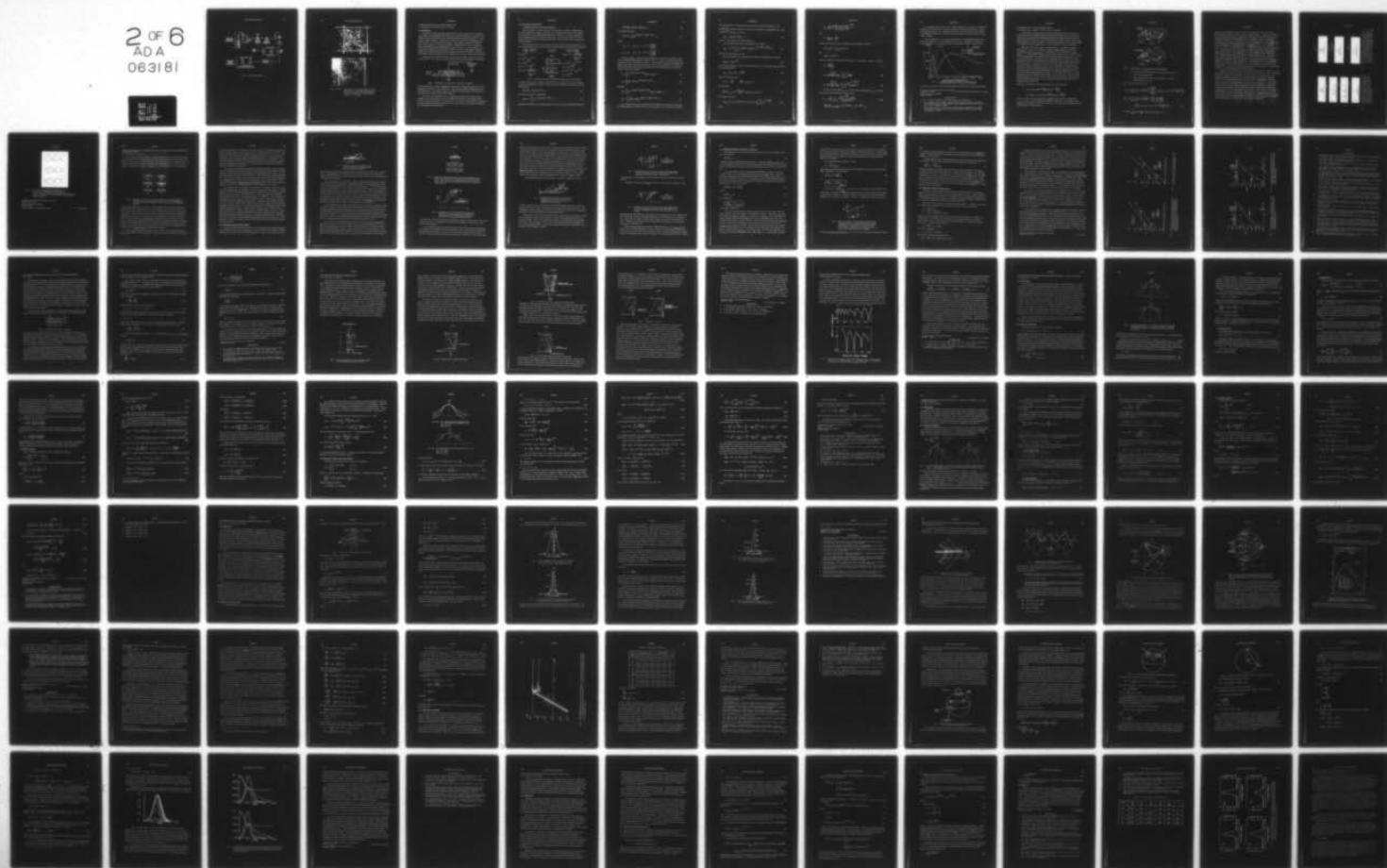
F44620-78-C-0074

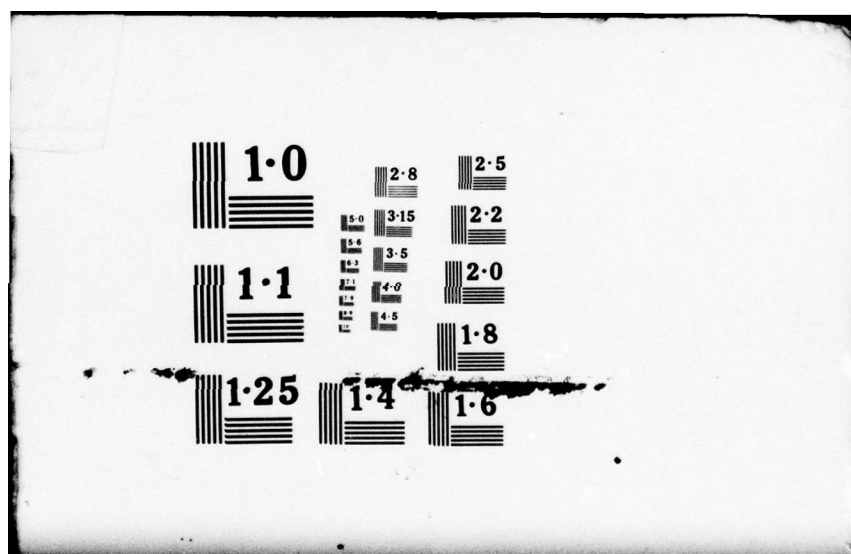
UNCLASSIFIED

POLY-MRI-452.43-78

NL

2 OF 6  
ADA  
063181





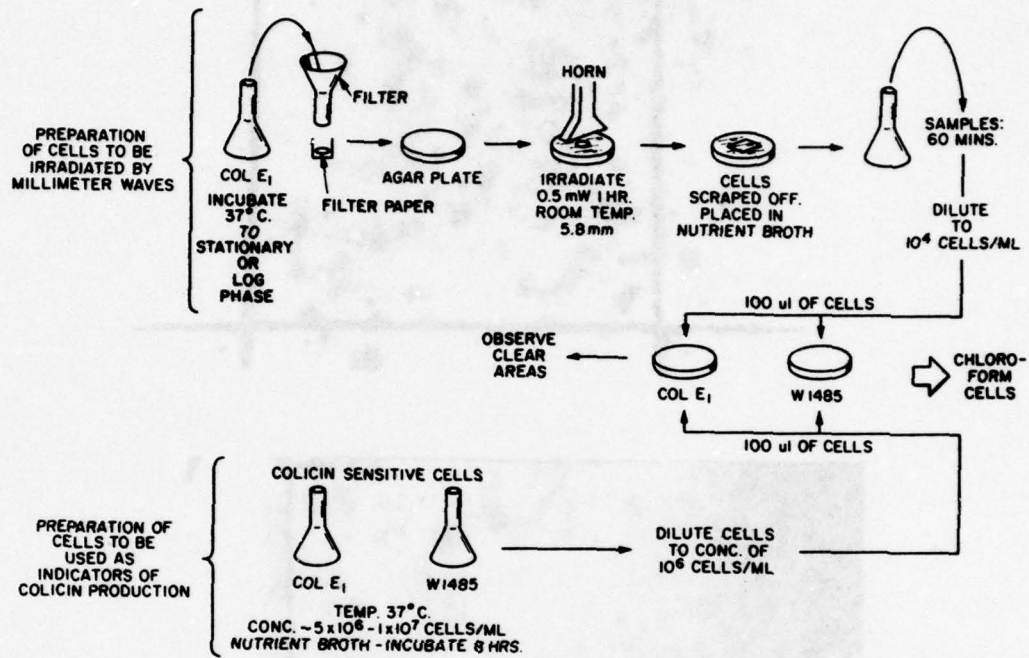


Fig. 9. Bacterial protocol.



Figs. 10 and 11. Dimensions of area under square =  $\text{cm}^2$ ; Light areas - clear areas where colicinogenic bacteria have destroyed indicator bacteria; Large areas = .8 mm; Small areas - .2 - .5 areas.

## ACoustoelectric REAL TIME CORRELATOR

L. Rosenheck, H. Schachter and W. C. Wang

## A. Introduction

It has been demonstrated experimentally that surface acoustic wave correlation of two signals can be performed by using space charge nonlinearity in a structure consisting of a thin semiconductor layer inserted between two different piezoelectric substrates;  $\text{LiNbO}_3$ -Si-BGO. The two surface waves are collinear and propagate in the same direction. Both the theoretical analysis and computer results are presented here. In the analysis, 1) the diffusion effect is included and 2) neither thin film nor thick semiconductor layer approximation is used, since  $d < \lambda_a$ ,  $d > \lambda_D$ , where  $d$ ,  $\lambda_a$  and  $\lambda_D$  are the respective semiconductor thickness, acoustic wave length and Debye length. The charge distribution and the nonlinear interaction strength as a function of semiconductor thickness (thin film case included) and resistivity will also be discussed.

In this paper we study the optimum acoustoelectric voltage across the structure shown in Fig. 1. This structure consists of two semi-infinite piezoelectric substrates,

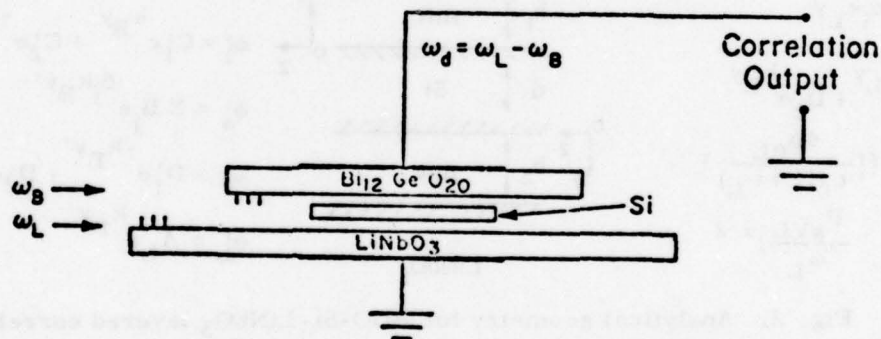


Fig. 1. Structure of acoustoelectric real time correlator.

one of  $\text{LiNbO}_3$  [010] cut [001] propagating and the other of  $\text{Bi}_{12}\text{GeO}_{20}$  [001] cut [110] propagating. Between the two substrates is a thin Si plate which is sandwiched in such a way that there is an air gap of distance  $h_1$  between the  $\text{Bi}_{12}\text{GeO}_{20}$  and the Si and one of  $h_2$  between the  $\text{LiNbO}_3$  and the semiconductor.

The role of the semiconductor is to couple two collinear surface waves which propagate in the same direction, one on the BGO and the other on the  $\text{LiNbO}_3$  surface via space charge nonlinearity.<sup>1,2</sup> This gives use to an open circuit voltage at a frequency equal to the difference between frequency of the acoustic wave traveling on the  $\text{LiNbO}_3$ ,  $\omega_2$ , and the BGO traveling wave  $\omega_3$ .

## B. Theoretical Consideration

### 1. Fields and Current in the Semiconductor

We will assume an input signal  $f(t) = 2A \cos(\omega_L t)$  at the  $\text{LiNbO}_3$  transducer and  $g(t) = 2B \cos(\omega_B t)$  at the BGO transducer. The two input signals will excite acoustic traveling waves in the  $\text{LiNbO}_3$  and the BGO with velocities of  $v$  and  $v_B$ , respectively. The accompanying electric fields will give rise, in a first order, to electric fields and charges in the silicon traveling with the same velocity as the electric fields in the  $\text{LiNbO}_3$  and BGO.

As depicted in Fig. 2 it is convenient for one to treat this layered structure as two separate problems and at a later time consider the nonlinear interaction. Following

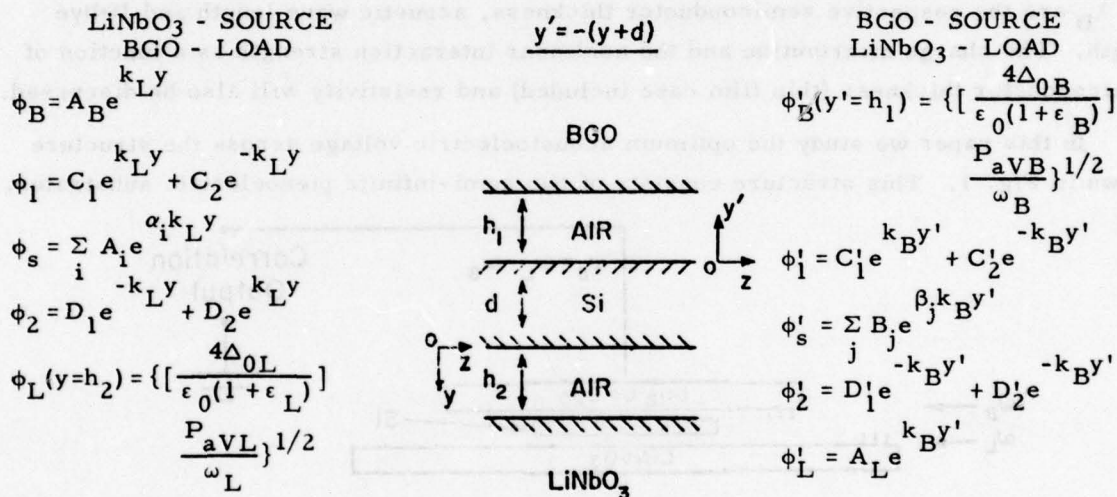


Fig. 2. Analytical geometry for BGO-Si- $\text{LiNbO}_3$  layered correlator.

a method similar to that of Lakin and Shaw<sup>3</sup> we first consider the case where  $\text{LiNbO}_3$  is the source and BGO is the load. This gives use to a differential equation within the semiconductor

$$\psi_s^{IV} - k_L^2(1 + \alpha^2) \psi_s^{II} + k_L^4 \alpha^2 \psi_s = 0 \quad (1)$$

The solution of which is the potential

$$\psi_s(t, y, z) = \sum_{i=1}^4 A_i e^{\alpha_i k_L y} e^{j(\omega_L t - k_L z)} + \text{c. c.} \quad (2)$$

Similarly for  $\text{LiNbO}_3$  as the load and BGO as the source we get a differential equation

$$\psi_s^{IV} - k_B^2(1 + \beta^2) \psi_s^{II} + k_B^4 \beta^2 \psi_s' = 0 \quad (3)$$

the solution of which is

$$\psi_s'(t, y, z) = \sum_{j=1}^4 B_j e^{-\beta_j k_B (y+d)} e^{j(\omega_B t - k_B z)} + c.c. \quad (4)$$

where

$$\alpha_{1,2} = \pm 1 \quad \beta_{1,2} = \pm 1 \quad \alpha^2 = 1 + \frac{\sigma + j\omega_L \epsilon_s}{\epsilon_s D k_L^2}$$

$$\alpha_{3,4} = \pm \gamma \quad \beta_{3,4} = \pm \beta \quad \beta^2 = 1 + \frac{\sigma + j\omega_B \epsilon_s}{\epsilon_s D k_B^2}$$

The  $A_{i,s}$  and  $B_{j,s}$  coefficients are determined by using the configuration shown in Fig. 2 and following boundary conditions:  $J_y(0) = 0$ ,  $J_y(-d) = 0$  the potential and displacement field are continuous at 0 and  $-d$ . We also note that  $\varphi_L(y = h_2)$  and  $\varphi_B'(y' = h_1)$  represent the source potential proportional to the acoustic input power. Associated with the potential  $\psi_s$  and  $\psi_s'$  one can easily find the corresponding electric field and charges in the semiconductor

$$E_L = \sum_{i=1}^4 A_i e^{\alpha_i k_L y} e^{j(\omega_L t - k_L z)} (jk_L \hat{z} - \alpha_i k_L \hat{y}) + c.c. \quad (5)$$

$$\rho_L = \sum_{i=1}^4 k_L^2 (1 - \alpha_i^2) A_i e^{\alpha_i k_L y} e^{j(\omega_L t - k_L z)} + c.c. \quad (6)$$

Similarly,

$$E_B = \sum_{j=1}^4 B_j e^{-\beta_j k_B (y+d)} e^{j(\omega_B t - k_B z)} (-jk_B \hat{z} + \beta_j k_B \hat{y}) + c.c. \quad (7)$$

$$\rho_B = \epsilon_s \sum_{j=1}^4 k_B^2 (1 - \beta_j^2) B_j e^{-\beta_j k_B (y+d)} e^{j(\omega_B t - k_B z)} + c.c. \quad (8)$$

The coupling of the electric fields due to the  $\text{LiNbO}_3$  SAW with the charges due to the SAW propagating on the BGO and vice versa will give rise to nonlinear currents in

the semiconductor. If the frequencies  $\omega_B$  and  $\omega_L$  are chosen such that  $k_L = k_B$  (or  $\frac{\omega_B}{v_B} = \frac{\omega_L}{v_L}$ ) then the nonlinear current will be stationary and independent of  $z$ . The nonlinear current  $J_{NL}$  is given by

$$J_{NL} = \mu(E_L \rho_B^* + E_B^* \rho_L) + c.c. \quad (9)$$

where  $\mu$  is the mobility of the semiconductor.

If we substitute Eqs. (5) to (8) into Eq. (9) we find that

$$J_{NL} = \epsilon_s \mu k^3 \sum_{i=1}^4 \sum_{j=1}^4 A_i B_j^* (\beta_j^* - \alpha_i) (\alpha_i \beta_j^* + 1) e^{(\alpha_i - \beta_j^*)ky} e^{-\beta_j^*kd} e^{j\omega_d t} \quad (10)$$

where  $k = k_L = k_B$  and  $\omega_d = \omega_L - \omega_B$ .

We can assume that the potential at the difference frequency  $\omega_d$  is given by

$$\Phi_d(y, t) = \varphi_d(y) e^{j\omega_d t} \quad (11)$$

and from the equation of current in the semiconductor in one dimension (because there is no variation with  $z$ )

$$J_{dy} = \sigma E_{dy} + J_{NL} - D_n \frac{\partial \rho_d}{\partial y} \quad (12)$$

and the continuity equation

$$\nabla \cdot J_d = - \frac{\partial \rho_d}{\partial t} \quad \left( \frac{\partial J_{dy}}{\partial y} - j\omega_d \epsilon_s \rho_d = 0 \right) \quad (13)$$

We finally get

$$\frac{d}{dy} \left( \sigma E_{y, \omega_d} - \epsilon_s D_n \frac{d^2 E_{y, \omega_d}}{dy^2} + j\omega_d \epsilon_s E_{y, \omega_d} + J_{NL} \right) = 0 \quad (14)$$

the solution of which is

$$E_{Ly, \omega_d} = E_0 + E_1 \sinh \gamma(y+d) + E_2 \sinh \gamma y + \sum_{i=1}^4 \sum_{j=1}^4 K_{ij} e^{(\alpha_i - \beta_j^*)ky} \quad (15)$$

where

$$K_{ij} = \frac{\mu k^3}{D_n} \frac{A_i B_j^* (\beta_j^* - \alpha_i) (1 + \alpha_i \beta_j^*) e^{-k \beta_j^* d}}{\{(\alpha_i - \beta_j)^2 k^2 - \gamma^2\}} \quad (16)$$

and

$$\gamma^2 = \frac{\sigma}{D_n \epsilon_s} (1 + j \frac{\omega_d}{\omega_c})$$

In order to solve for  $E_0$ ,  $E_1$ , and  $E_2$  we use the boundary conditions

$$\epsilon_s E_{y, \omega_d}(0) = \epsilon_s E_{y, \omega_d}(-d) = \epsilon_0 E$$

and

$$J_{yd}(0) = J_{yd}(-d) = 0$$

where  $E$  is the electric field at the frequency  $\omega_d$  in the adjacent medium. We obtain

$$E_0 = \frac{j \omega_d \epsilon_0 E}{\sigma + j \omega_d \epsilon_s} \quad (17)$$

$$E_1 = \frac{\frac{\epsilon_0}{\epsilon_s} E \omega_c}{\sinh \gamma d (\omega_c + j \omega_d)} - \sum_{i=1}^4 \sum_{j=1}^4 \frac{K_{ij}}{\sinh \gamma d} \quad (18)$$

$$E_2 = \frac{\frac{\epsilon_0}{\epsilon_s} E \omega_c}{\sinh \gamma d (\omega_c + j \omega_c)} + \sum_{i=1}^4 \sum_{j=1}^4 \frac{K_{ij} e^{-(\alpha_i - \beta_j^*) k d}}{\sinh \gamma d} \quad (19)$$

The open circuit voltage ( $V_{o.c.}$ ) is then found by integrating  $E_{y, \omega_d}$  across the semiconductor and retaining the nonlinear part.

$$V_{o.c.} = \frac{\mu k^2}{D_n} \sum_{i=1}^4 \sum_{j=1}^4 \frac{A_i B_j^* e^{-k \beta_j^* d} (1 + \alpha_i \beta_j^*)}{(\alpha_i - \beta_j^*)^2 k^2 - \gamma^2} \quad (20)$$

$$\cdot \left\{ \frac{(\beta_j^* - \alpha_i) k}{\gamma} \tanh \frac{\gamma d}{2} \left[ 1 + e^{-(\alpha_i - \beta_j^*) k d} \right] + e^{-(\alpha_i - \beta_j^*) k d} - 1 \right\}$$

It is apparent that the open circuit voltage should have an optimum at a particular frequency for a particular semiconductor thickness. If the semiconductor is "thick" we would expect little or no coupling between the electric field and the space charges of the  $\text{LiNbO}_3$  and the BGO. Similarly, if the semiconductor is thin the area of interaction is small.

If we consider Fig. 3, we see that indeed there is an optimum thickness  $d$  for a particular frequency. It is also apparent that if one was interested in an optimum voltage regardless of the frequency or the semiconductor thickness it would only be

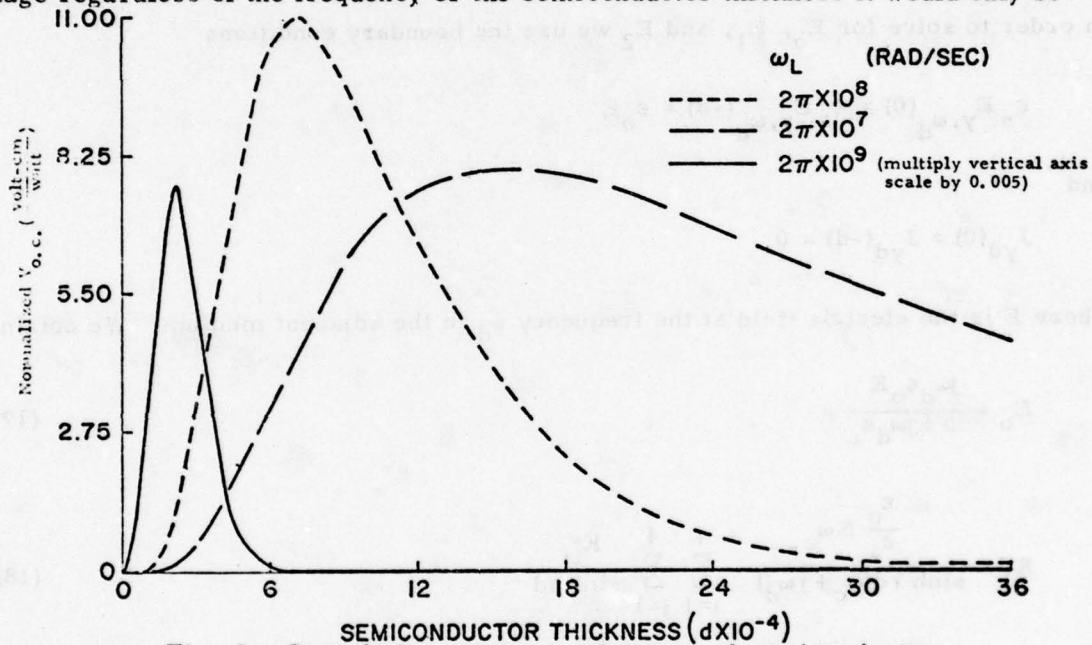


Fig. 3. Correlation output as a function of semiconductor thickness ( $d$ ) with frequency ( $\omega_L$ ) as a parameter.

necessary to vary the thickness and the frequency until  $V_{oc}$  was an absolute maximum.

Joint Services Technical Advisory Committee

F44620-74-C-0056

L. Rosenheck, H. Schachter and W. C. Wang

National Science Foundation

ENG 76-18901

#### REFERENCES

1. W. C. Wang, IEEE Ultrasonic Symposium, New York (1971).
2. W. C. Wang and K. S. Meng, "Surface-Wave Correlation and Time Scaling in a Structure of  $\text{Bi}_{12}\text{GeO}_{20}$ - $\text{SiLiNbO}_3$ ," Appl. Physics Letters, Vol. 27, No. 7, pp. 375-377 (October 1, 1975).
3. K. M. Lakin and H. J. Shaw, "Surface Wave Delay Line Amplifiers," IEEE Transactions on Microwave Theory and Techniques, MTT-17, pp. 912-920 (November 11, 1969).

## AN ACOUSTOELECTRIC FM DEMODULATOR

H. Schachter, W. C. Wang, F. Cassara and L. Rosenheck

Acoustoelectric signal processors such as the convolver and correlator are built based on (i) strong space charge nonlinearity induced by SAWs in an adjacent semiconductor and (ii) the filtering action generated by spatial integration over the length of nonlinear interaction. Utilizing these special features again, a new type of FM demodulator has been developed. This type of demodulator is structurally simple, Si on  $\text{LiNbO}_3$  and can be operated at a carrier frequency in excess of 100 Mhz.

Figure 1(a) describes the basic device geometry. A frequency modulated signal,  $f(t) = A \cos(\omega_c t + \Delta\omega \int^t g(\tau) d\tau)$  is applied simultaneously to the pair of interdigital transducers  $T_1$  and  $T_2$ . The transducers are separated by a small distance  $\ell_0$ . The applied FM signal will induce two collinear surface acoustic waves which are propagating in the same direction and with a time separation of  $t_0 = \ell_0 / v$ .  $v$  is the SAW velocity. When the two waves propagate under semiconductor, the piezoelectric field associated with the waves will induce space charge waves inside the semiconductor. Through the nonlinear interactions among the induced space charge waves and the piezoelectric field waves, many higher order signals are generated. However, after integrating along the semiconductor length  $L$ , only the term which performs the FM demodulation is of significant amplitude and can be detected across the semiconductor terminals. For simplicity and clarity, we will examine the case of a single tone FM signal; i. e.,  $g(\tau) = \cos \omega_m \tau$ . The sum of the input signal and its delay squared,  $\{f(t) + f(t-t_0)\}^2$  will give rise to several second order nonlinear mixing terms. It can be shown that the only term which is not negligible and of interest is:

$$\begin{aligned} 2f(t)f(t-t_0) &= 2AB \cos\{\omega_c t + \beta \sin \omega_m t\} \cos\{\omega_c(t-t_0) + \beta \sin \omega_m(t-t_0)\} \\ &= AB \cos\{\omega_c t_0 + [2\beta \sin \frac{\omega_m t_0}{2}]\cos(\omega_m t - \frac{\omega_m t_0}{2})\} \\ &\quad + AB \cos\{2\omega_c t - \omega_c t_0 + [2\beta \cos \frac{\omega_m t_0}{2}]\sin(\omega_m t - \frac{\omega_m t_0}{2})\} \end{aligned} \quad (1)$$

where  $A$  and  $B$  are associated with the amplitude of  $f(t)$  and  $f(t-t_0)$  respectively.  $\beta$ ,  $\omega_m$ ,  $\Delta\omega$  and  $\omega_c$  are the respective modulation index ( $\Delta\omega/\omega_m$ ), modulation rate, peak frequency deviation and carrier frequency. The open circuit voltage across the semiconductor is then of the form

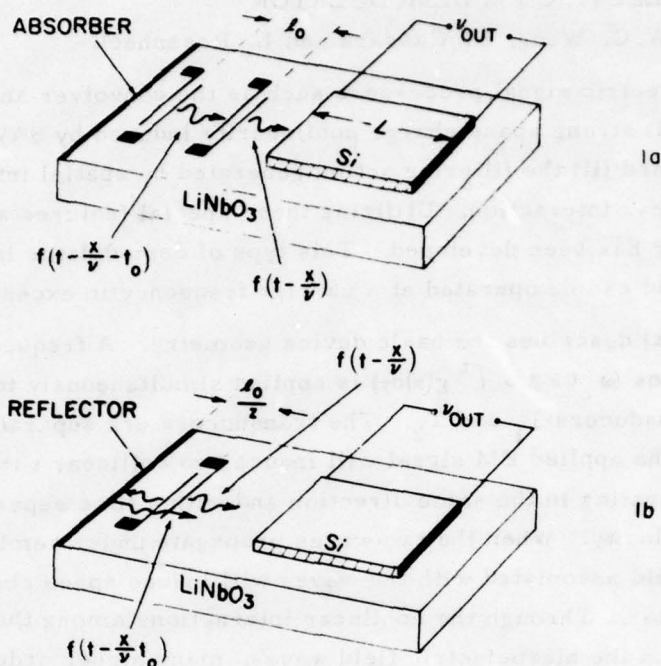


Fig. 1. Geometries of Acoustoelectric FM Demodulator.

a - The delay time  $\frac{l_0}{v_s}$  introduced by the spatial separation of the two input transducers.

b - The delay time  $\frac{l_0}{v}$  introduced by reflection. A single input transducer is used.

$$\begin{aligned}
 v_{o.c.} &= \int_0^L K_2 2f(t - \frac{x}{v}) f(t - \frac{x}{v} - t_0) dx \\
 &= K_2 LAB \sum_{n=-\infty}^{\infty} J_n \left[ 2\beta \sin\left(\frac{\omega_m t_0}{2}\right) \right] \frac{\sin\left(\frac{n\omega_m L}{2v}\right)}{\left(\frac{n\omega_m L}{2v}\right)} \cos \left\{ n\omega_m t + \omega_c t_0 + \frac{n\pi}{2} - \frac{n\omega_m(L + \ell_0)}{2v} \right\} \\
 &+ K_2 LAB \sum_{p=-\infty}^{\infty} J_p \left[ 2\beta \cos\left(\frac{\omega_m t_0}{2}\right) \right] \frac{\sin\left(\frac{(2\omega_c + p\omega_m)L}{2v}\right)}{(2\omega_c + p\omega_m) \frac{L}{2v}} \\
 &\cos \left\{ (2\omega_c + p\omega_m)t + \omega_c t_0 - \frac{\omega_c L}{v} - \frac{p\omega_m(L + \ell_0)}{2v} \right\} \\
 &\approx K_2 LAB \sum_{n=-\infty}^{\infty} J_n(\Delta\omega t_0) \left\{ \cos n\omega_m t + \omega_c t_0 + \varphi \right\}, \text{ for } n \leq 2.
 \end{aligned} \tag{2}$$

where  $K_2$  is a constant in terms of the 2nd order nonlinear coefficient and other device parameters such as carrier density, mobility and etc.  $J_n(\Delta\omega t_0)$  is the Bessel function of  $n$ th integer order. Equation (2) reveals that the novel SAW device does indeed demodulate the received FM signal ( $n = \pm 1$  terms). All other terms are negligible provided  $\Delta\omega t_0 < 1$  since  $J_n(\Delta\omega t_0) \ll J_1(\Delta\omega t_0)$  for  $n \geq 2$ . It is noted in the above expression that the action of spatial filtering comes in the form of  $\frac{\sin \theta}{\theta}$ . As a numerical example, let us consider the case when  $\frac{L}{v} \approx 3 \mu\text{sec}$ ,  $\omega_m \approx 8\pi \times 10^3$  rad./sec. and  $\omega_c \approx 2\pi \times 10^8$  rad./sec. The  $\frac{n}{\theta}$  factor ( $\theta_n = n\omega_m L/2v$ , associated with  $J_n$ ) approaches to 1 for  $n \leq 2$ , and the factor  $\sin \theta_p / \theta_p$  (associated with  $J_p$ ) approaches to zero for any  $p$ . Therefore, the only term remaining in the open circuit voltage expression is that related with the modulation signal. Further, it is noted that  $2\beta \sin(\omega_m t_0/2) = \Delta\omega t_0$ , since  $\omega_m t_0/2 \ll 1$ . For linearity,  $\Delta\omega t_0$  is required to be less than one. For example, if a frequency deviation,  $\Delta f = 20$  Mhz is needed, one requires a delay to  $t_0 \leq 5$  n sec. and a carrier frequency  $f_c \approx 10^3$  Mhz. A unique feature inherent in this device is that one can adjust  $\omega_c t_0$  to select a set of desired order, (even, odd or both) of Bessel functions to operate with. Experimental results are in agreement with the derived expression.

A typical set of experiments are described here in which  $t_0$  is set at  $0.34 \mu\text{sec}$ . The oscillograms in Fig. 2 are taken at constant carrier frequency, 95 Mhz at varying deviations. (At this carrier frequency only odd harmonics are observed.) The observed disappearances (first zeros) of fundamental, third harmonic and fifth harmonic are corresponding to  $\Delta\omega t_0 = 3.74$ , 6.4 and 8.8 respectively. They are compared with theoretical predictions of  $\Delta\omega t_0 = 3.81$ , 6.38 and 8.77. The oscillograms in Fig. 3 are taken under conditions of constant frequency deviation, but varying carrier frequencies. In oscillogram 3(a) only the odd (fundamental) is predominant, in 3(b) both even and odd are present, in 3(c) only the even (second) is predominant. From 3(a) to 3(c) the change in  $\omega_c t_0$  is  $(\omega_c t_0)_a - (\omega_c t_0)_c = 0.54\pi$ . Figure 4 gives the comparison between the shapes of the input modulation signal  $g(\tau)$  and that of the device's output demodulated signal. Clearly successful FM demodulation is achieved even for the case of nonsinusoidal modulation.

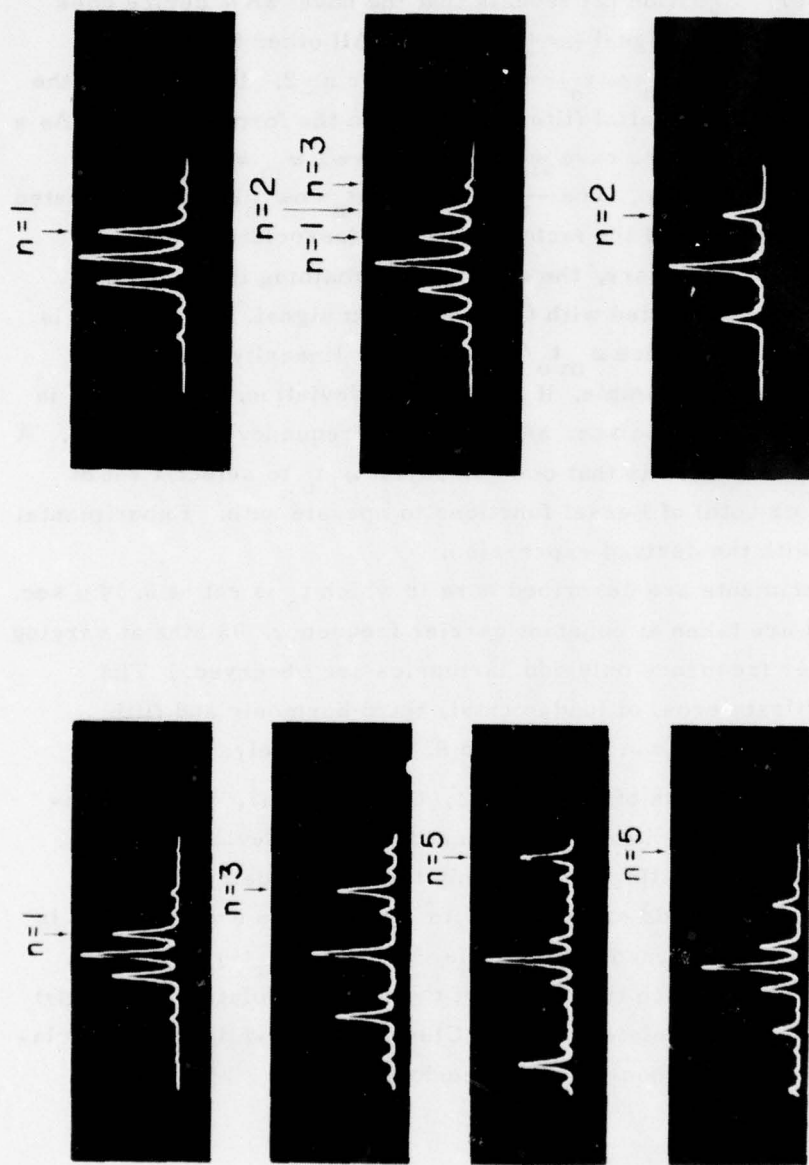


Fig. 2. Observed harmonic contents as a function of  $\Delta\omega t_0$ . Carrier frequency  $f_c = 95$  MHz,  $t_0 = 0.34 \mu\text{sec}$ . Modulation rate  $f_m = 2.2$  kHz.  $\omega_c t_0$  is set for negligible even harmonic. The center trace represents local oscillator frequency (500 kHz).

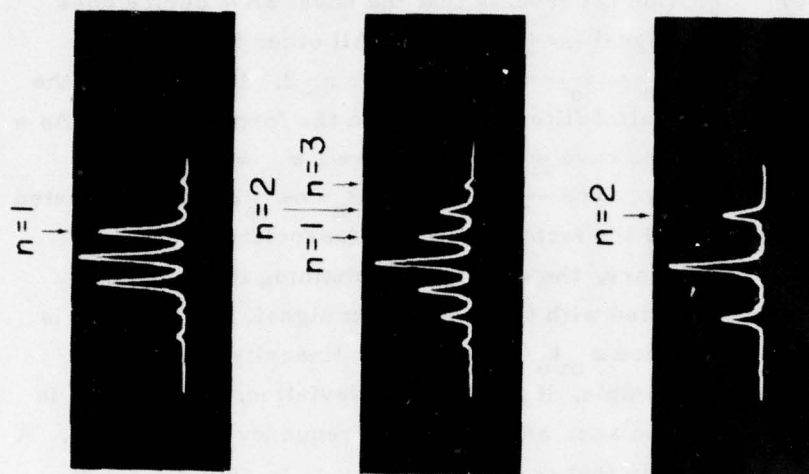


Fig. 3. Observed harmonic contents as a function of  $f_c$ . Constant peak frequency deviation  $\Delta\omega = 7 \times 10^6$  rad./sec.,  $t_0 = 0.34 \mu\text{sec}$ . In (a), only fundamental (odd) is predominant. In (b), both even and odd harmonics are present. In (c), only 2nd harmonic (even) is predominant. The change  $(\omega_c t_0)_a - (\omega_c t_0)_b = 0.54 \pi$ .

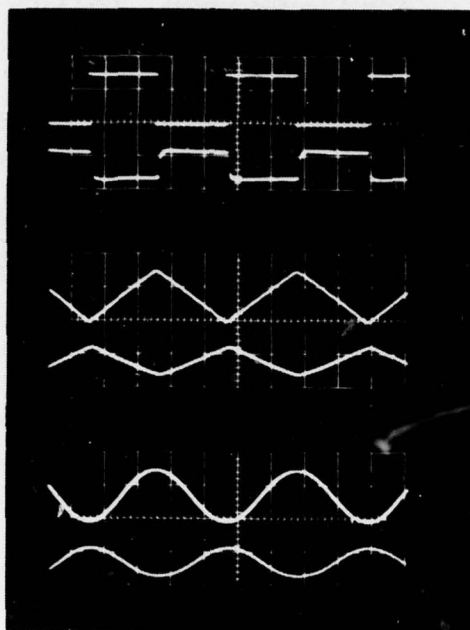


Fig. 4. Top trace, input modulation signal  $g(t)$   
Lower trace, demodulated output FM signal.  
(the  $180^\circ$  phase inversion between input and output can be  
made zero degrees by interchanging the semiconductor  
output terminals).

Joint Services Technical Advisory Committee  
F44620-74-C-0056

National Science Foundation  
ENG 76-18901

Air Force Office of Scientific Research

H. Schachter

# A SIMPLE CRITERION FOR PREDICTING LEAKY WAVES ON RIB WAVEGUIDES FOR INTEGRATED OPTICS

S. T. Peng and A. A. Oliner

Three-dimensional waveguides for integrated optics can be divided into two broad classes: those that consist of a dielectric strip which is deposited on a dielectric substrate or else a channel which is diffused or ion-implanted into it, as shown in Fig. 1(a) and 1(b), and those which utilize an additional thin dielectric layer on top of the substrate, as illustrated in Figs. 1(c) through 1(f). The waveguides in the first class are

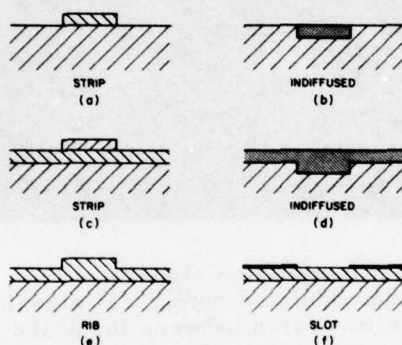


Fig. 1. Examples of three-dimensional waveguides for integrated optics. Waveguides (a) and (b) fall into the class for which no leakage is possible; the rest are members of the class for which some modes may leak.

essentially modifications of a dielectric rod waveguide; if the strip were absent, no surface wave could be guided at all. For the waveguides in the second class, the additional thin layer would guide a wide surface wave beam in the absence of the guiding strip region. The two classes thus differ from each other because of the added thin dielectric layer, or film. The significance of this distinction in connection with this report is that all modes propagating along waveguides in the first class are purely guided, with no leakage possible (assuming that the waveguides are uniform -- leakage can occur if the waveguiding structure is tapered or periodically-modulated along its length). On the other hand, as we shall see below, some modes on waveguides of the second class are leaky modes under appropriate conditions.

It is not generally known that leaky modes can exist on the large class of waveguides denoted above as the second class. In that class, the thin dielectric film serves

to confine the guided wave laterally to the region of the strip. Examples of such waveguides include a dielectric strip with refractive index less than that of the film,<sup>1</sup> index greater than that of the film,<sup>2</sup> and equal to that of the film,<sup>3-5</sup> when it is called a "rib" waveguide. Other structures employ ion-implanted layers,<sup>6,7</sup> or metallic coatings everywhere but in the guiding region, and are known as "slot" waveguides.<sup>8-11</sup> This class of waveguides has also been generically referred to as "slab-coupled" guides.<sup>12</sup> All of these guiding structures have been built and tested in the context of directional couplers, modulators or switches.

In an earlier publication,<sup>5</sup> we presented a qualitative explanation for the leakage effects, an underlying analysis for an approximate evaluation of this leakage, and some numerical results for the rib waveguide for some specific parameter values. We are currently performing a more accurate analysis, using a rigorous phrasing of the problem. In this report, however, we are concerned with how to predict when leakage will occur without having to first solve the problem (which is quite complicated). In this connection we present a simple but accurate criterion which is based on the known dispersion curves for the constituent portions of the waveguide's cross section.

It is important for two reasons to know whether or not leakage is present. Since these optical waveguides are intended for use in an integrated circuit fashion, unwanted leakage can cause cross talk between neighboring components and thus deteriorate the performance of the circuit. On the other hand, novel components can be designed which make deliberate use of the leakage present. An example of such a component for integrated optics is a novel leaky wave directional coupler.<sup>13</sup> This coupler consists of the usual two strip waveguides located parallel to each other and spaced a certain distance apart, but here the spacing is made so great that the coupling would be negligible if leakage were not present. It is found that the coupling is not sharply sensitive to the separation between the strips, a feature which may have practical import. Of even greater interest is the coupler's potential use as a mode stripper or purifier. In general, when one wants to excite the TE mode on the optical strip guide some amount of TM mode energy will inevitably be present. Since, as will be shown below, the lowest TM mode on the optical strip waveguide leaks and the lowest TE mode does not, such a leaky wave coupler can be used to tap away all the TM mode energy without disturbing the TE mode at all.

#### A. The Physical Basis for the Leakage

A typical rib waveguide is shown in Fig. 2; the material of the strip is the same as that of the thin film. The strip produces a central region with an effective refractive index greater than that of the surrounding regions; as a result, the surface wave that

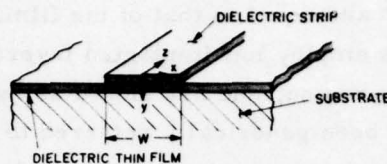


Fig. 2. Geometry of the rib waveguide. The central strip and the surrounding thin film are of the same material.

is present on the thin film is pulled in by the strip region and confined to its neighborhood. The guiding process is usually viewed in terms of waves in the strip region bouncing back and forth at an angle between the two sides of the strip, undergoing total reflection at each bounce.

The cross section is considered to consist of two constituent regions, the inner strip region and the outer portions. Theoretical analyses for the propagation characteristics of strip waveguides have appeared in the literature,<sup>1,2,12,14</sup> but these analyses assume the presence of only one mode type, TE or TM, in each of the constituent regions comprising the cross section of the strip waveguide. When the field behavior at the strip sides is viewed more carefully, however, it is easy to see that a TE or TM mode incident on a strip side produces not only a reflected and transmitted wave of its own type, but also excites a reflected and a transmitted wave of the other type, plus a continuous spectrum (which is purely reactive here). This coupling at the strip sides between TE and TM modes produces the leakage effects discussed here.

If we assume that the dielectric film thickness is such that the constituent regions in the cross section can each support only one TE mode and one TM mode (which corresponds to the usual range of operation), then a transverse equivalent network for the rib waveguide which takes into account the TE-TM coupling mentioned above has the form seen in Figure 3. The transmission line portions represent the independent modes in the cross section, and the step junction discontinuities which correspond to the strip sides are represented by the boxes which couple the transmission lines.

Let us next consider separately each of the constituent regions comprising the guide cross section, as if each were infinitely wide. The dispersion curves for these regions are presented in Fig. 4 for both the TE and TM modes, in the form of the effective refractive index  $n_{\text{eff}} (= \beta/k = \lambda/\lambda_g)$  as a function of  $t/\lambda$ , where thickness  $t$  is defined in the inset drawing. The TE mode is seen to be the dominant mode, in the sense that it is the slower mode (higher value of  $n_{\text{eff}}$ ) and it has the smaller cutoff frequency

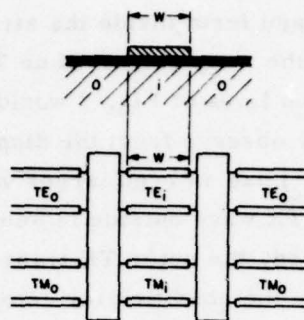


Fig. 3. Transverse equivalent network for the rib waveguide which takes into account the TE-TM coupling produced at the strip sides. The symbols *i* and *o* signify respectively the inner (strip) and outer (film) constituent portions of the waveguide cross section.

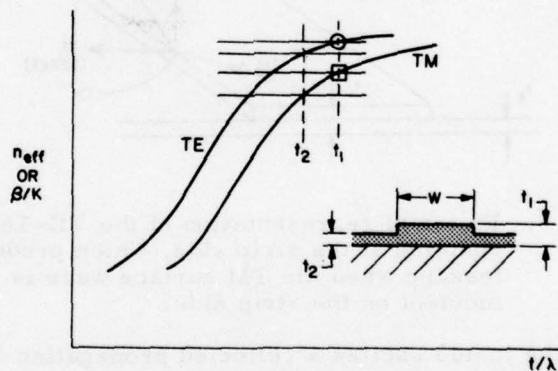


Fig. 4. Dispersion curves for the basic surface waves which exist on the constituent (strip and film) portions of the rib waveguide cross section, if each portion is considered infinitely wide.

(or thickness). The curve for the other polarization (TM) is located nearby. The vertical line  $t_1$  in Fig. 3 corresponds to the (thicker) strip region, and the line  $t_2$  to the outside (film) regions.

For complete guiding, the waves bouncing back and forth inside the strip region at an angle to the strip sides must be propagating, and the waves present outside must be transversely evanescent. In terms of the transverse equivalent network of Fig. 3, the inner TE and TM transmission lines must be above cutoff and the outer TE and TM

transmission lines must be below cutoff. Suppose that the waveguide is excited with the electric field oriented in the vertical ( $y$ ) direction, so that the basic excitation is TM. As this TM wave bounces back and forth inside the strip region, it couples some TE energy at each reflection from the strip sides. If no TE energy were coupled, then the inner and outer TM transmission lines of Fig. 3 would clearly be above and below cutoff, respectively. But, we may observe from the dispersion curves in Fig. 4 that the TE wave outside (at thickness  $t_2$ ) has an even larger value of  $n_{\text{eff}}$  than does the TM wave inside (at  $t_1$ ), so that the TE wave outside is seen to be propagating rather than evanescent; alternatively phrased, the outer TE transmission line is above cutoff rather than below cutoff. Hence the condition for complete guiding is not satisfied, and leakage of energy transversely away from the waveguide is being produced. On the other hand, if the initial waveguide excitation were of the TE type, a corresponding inquiry would indicate that the resulting guided wave would be completely bound.

The leakage produced when TM excitation is incident is depicted in Figure 5. A

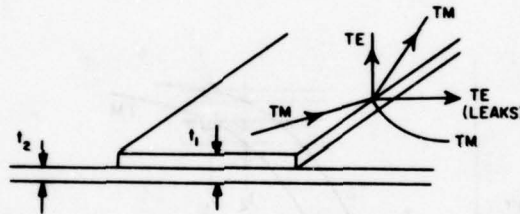


Fig. 5. Pictorial representation of the TE-TM coupling at the strip side, which produces leakage when the TM surface wave is incident on the strip side.

TM wave incident from the inside excites a reflected propagating TM wave inside, an evanescent TM wave outside, and propagating TE waves both outside and inside. The propagating TE wave outside carries away a small amount of energy at each reflection, resulting in a leaky mode rather than a purely bound mode.

Depending on how close the dispersion curves of Fig. 4 are to each other, the TE wave outside may or may not be evanescent, and the resulting guided mode may or may not be leaky. Thus, some modes may be leaky or not depending on the geometric parameters of the waveguide. We have here indicated qualitatively how leakage can occur, but we present below a criterion for leakage which permits a quantitative determination of when leakage is present.

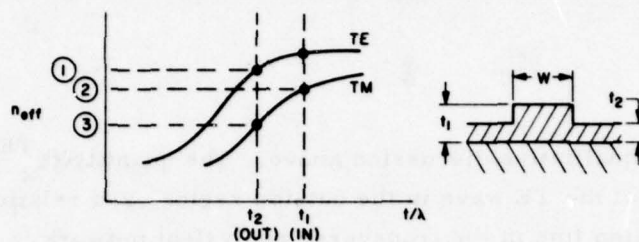


Fig. 6. Dispersion curves for the first class of rib waveguides, for which value ① lies above values ② and ③, so that leakage occurs for all values of strip width.

unconditionally, so that leakage in that mode will be present for all values of strip width  $W$ .

Dispersion curves for the second of the two cases are given in Figure 7. The

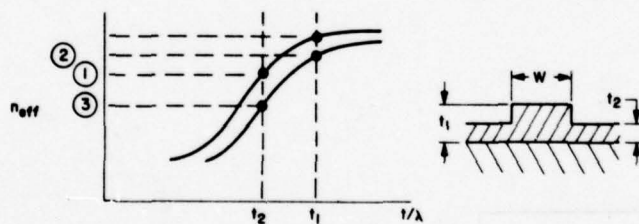


Fig. 7. Dispersion curves for the second class of rib waveguides, for which value ① lies between values ② and ③. For this case, leakage will occur only for sufficiently narrow strips.

designations ①, ② and ③ mean the same thing here as in Fig. 6, but we observe here that value ① lies between values ② and ③ and not above them both, as in the first case. Hence we conclude that leakage will no longer occur for all strip widths,  $W$ . In order to determine for what values of  $W$  leakage will be present, we must solve for  $(n_{\text{eff}})_{\text{guide}}$  as a function of  $W$ .

Thus, inequality ④ is not helpful as it stands because  $k_z$  (or  $(n_{\text{eff}})_{\text{guide}}$ ) is not known until the complete problem is solved. However, we may employ an approximation to  $k_z$  to provide us with a useful criterion. The so-called equivalent index method,<sup>1</sup> which is a modification of the Marcatili approach,<sup>15</sup> or the equivalent dielectric constant

### B. A Simple Quantitative Criterion for Leakage

With reference to the coordinate system indicated on Fig. 2, the explicit condition for leakage is

$$(k_x^{\text{TE}})_{\text{out}}^2 > 0, \quad (1)$$

consistent with the qualitative discussion above. The quantity  $(k_x^{\text{TE}})_{\text{out}}$  is the transverse wavenumber of the TE wave in the outside region, and relation (1) states that the outer TE transmission line in the transverse equivalent network in Fig. 3 is above cut-off.

The actual TE wave outside, with wavenumber  $(k_s^{\text{TE}})_{\text{out}}$ , where  $s$  represents "surface wave," will be propagating away from the strip at some angle, and it will possess components  $k_z$  along the waveguide direction and  $(k_x^{\text{TE}})_{\text{out}}$  perpendicular to it in the plane of the thin dielectric film. Quantity  $k_z$  needs no other qualifying indices because all constituents of the net guided mode along the waveguide must possess the same  $k_z$  value. We may thus write

$$(k_s^{\text{TE}})_{\text{out}}^2 = k_z^2 + (k_x^{\text{TE}})_{\text{out}}^2 \quad (2)$$

so that

$$(k_s^{\text{TE}})_{\text{out}} > k_z \quad (3)$$

or, dividing by  $k$ ,

$$(n_{\text{eff}}^{\text{TE}})_{\text{out}} > (n_{\text{eff}})_{\text{guide}} \quad (4)$$

in the light of relation (1).

The range of possible rib waveguide geometries falls into two cases. The dispersion curves for the first of these cases appear in Figure 6. The value of  $n_{\text{eff}}$  indicated as ① corresponds to  $(n_{\text{eff}}^{\text{TE}})_{\text{out}}$ . Furthermore,  $(n_{\text{eff}})_{\text{guide}}$  will vary between the values denoted by ② and ③, although it may not follow the TM curve itself. The validity of that statement is easily verified. When the strip width  $W$  becomes extremely large, the inside region of thickness  $t_1$  dominates the geometry, and the value of  $(n_{\text{eff}})_{\text{guide}}$  must approach value ②. (Recall that we are discussing the case of TM excitation.) When  $W$  becomes zero, the outside region of thickness  $t_2$  is present everywhere, and the value of  $(n_{\text{eff}})_{\text{guide}}$  must reduce to ③. For this case, therefore, we observe that value ① is higher than values ② and ③, i.e., condition ④ is satisfied

method,<sup>16, 17</sup> which is the first step of a transverse resonance procedure, can supply us with an approximation which is reasonably accurate and yet simple. In both of these approximate procedures, the TE-TM coupling is neglected entirely; as a result, neither method can predict any leakage, but the  $k_z$  values obtained are reasonably reliable under most conditions.

As an approximation, therefore, we take

$$k_z \approx (k_z^{\text{TM}})_{\text{approx.}} \quad (5)$$

where the  $k_z$  found will be purely real and corresponds to the case of TM waves alone. Then, using Eqs. (3) or (4) we write

$$(k_s^{\text{TE}})_{\text{out}} > (k_z^{\text{TM}})_{\text{approx.}} \quad (6)$$

or

$$(n_{\text{eff}}^{\text{TE}})_{\text{out}} > (n_{\text{eff}}^{\text{TM}})_{\text{guide approx.}} \quad (7)$$

Now, a plot of  $(n_{\text{eff}}^{\text{TM}})_{\text{guide approx.}}$  as a function of strip width  $W$  can be obtained relatively easily, avoiding the complications introduced by TE-TM coupling and the continuous spectrum contributions. A typical such plot for the dispersion curve features exhibited in Fig. 7 is shown in Figure 8. As seen, consistent with condition (7), the rib

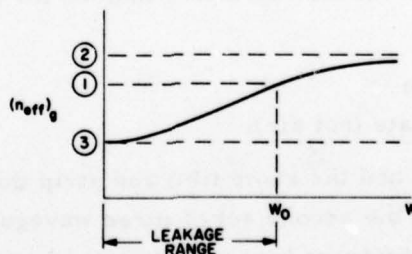


Fig. 8. Plot, as a function of strip width, of the effective refractive index for the mode guided along the rib waveguide for TM excitation, for the case considered in Fig. 7, where the effective index values are calculated approximately. Leakage is seen to occur only for the narrower strip widths.

waveguide will leak only for  $W < W_0$ , where  $W_0$  is the crossing point indicated in Figure

8. Leakage for TM excitation is thus seen to occur for this case only for small strip widths. It is therefore possible to design the waveguide geometry to avoid leakage altogether if one so wishes.

If TE excitation, rather than TM excitation, is incident on the rib waveguide, then the condition corresponding to relation (7) for TM excitation would become

$$(n_{\text{eff}}^{\text{TM}})_{\text{out}} > (n_{\text{eff}}^{\text{TE}})_{\text{guide}} \quad (8)$$

If this inequality were satisfied, then the TE guided mode would leak via the energy coupled at the strip sides to the TM wave. In terms of the dispersion plot of the type shown in Fig. 6 or Fig. 7, value ① would now correspond to  $(n_{\text{eff}}^{\text{TM}})_{\text{out}}$ , and values ② and ③ would be those on the TE curve at thicknesses  $t_1$  and  $t_2$ , respectively. Under those conditions, we observe that ① would always be lower than ③, so that condition (8) would never be satisfied. Thus, since the TE surface wave is the dominant mode in these dispersion plots, the TE mode on the rib waveguide would never leak, no matter what the value of strip width.

### C. Application to Rib Waveguide Devices

The rib waveguide structure was devised by the Bell Laboratories and then employed by them in several devices, such as modulators and switches. In their measurements, they did not detect the presence of leakage from the waveguides they employed. In an attempt to verify their results, we applied the criteria for leakage presented above to several rib waveguides which they used in their devices.<sup>18</sup>

All of their measurements were taken at the wavelength  $\lambda = 0.8966\mu$ . Six different waveguides were measured, but the refractive indices for all were the same;

$$\begin{aligned} n_f &= 3.4381, \text{ film} \\ n_s &= 3.3842, \text{ substrate} \\ n &= 1.720, \text{ superstrate (not air).} \end{aligned}$$

One set of three waveguides had the same film and strip thicknesses (or heights), and differed only in strip width; the second set of three waveguides had the same strip widths as the first set, but different film and strip thicknesses. Thus, using the notation in Fig. 2, we have for the first set

$$t_1 = 0.800\mu, \quad t_2 = 0.730\mu$$

and for the second set

$$t_1 = 0.700\mu, \quad t_2 = 0.630\mu.$$

For each set, the three strip widths were

$$W_1 = 3.50\mu, \quad W_2 = 3.00\mu, \quad W_3 = 2.50\mu.$$

Before we can apply the criteria for leakage developed here, we must have available the dispersion curves for the TE and TM modes in the strip and film regions separately, as if each region were infinitely wide. We thus would have one pair of curves for the first set of strip and film thicknesses, and another pair for the second set, as shown respectively in Figures 9 and 10. On these curves we have indicated the values ①, ② and ③, corresponding to those in Figures 6 and 7. We observe immediately that in both Figs. 9 and 10 value ① lies between values ② and ③, as in Fig. 7, indicating that the TM mode on these rib waveguides will leak only for sufficiently narrow strip widths, but not for all strip widths.

The next step is to determine the range of strip widths for which leakage will occur. Toward this end, so that criterion (7) can be used, we have calculated  $(n_{\text{eff}}^{\text{TM}})_{\text{guide}}$  as a function of strip width  $W$  for each set of thicknesses (corresponding to Figs. 9 and 10) using the approximate equivalent dielectric constant method, which is a simple procedure corresponding to the first step in a transverse resonance approach which neglects the TE-TM coupling and the continuous spectrum contributions. These approximate calculations are presented in Figs. 11 and 12, where the curves shown represent the two lowest modes, due to TM excitation, guided by the rib waveguide.

The range of  $W$  over which leakage occurs is indicated in Figs. 11 and 12 by the dashed lines, and  $W_0$  is the largest width for which leakage can be present. It is seen that for five of the six waveguides the strip width is larger than  $W_0$ , so that leakage would not be expected. For the sixth waveguide, the value of  $W_3$  occurs very close to  $W_0$ , right at the edge of the leakage range. Since the rate of leakage must come to zero at the edge of the range, the leakage rate for that waveguide must necessarily be quite small -- difficult to detect experimentally and probably negligible with respect to any practical consequence.

If the strip width were made narrower, leakage would then be possible. However, this criterion indicates only whether or not leakage will occur in principle; it says nothing about the amount of leakage. Depending on the geometrical parameters, the amount of leakage may be significant, or so small as to be of no practical import. The effective discontinuity provided by the sides of the strip is the key feature, because that is where the coupling occurs between the TE and TM modes. If the effective coupling is too small, the amount of leakage will be miniscule and the effect unimportant. However, it is necessary to solve the complete boundary value problem, including the TE-TM coupling, before one can determine the leakage rate quantitatively -- a short cut of the type presented here cannot be developed.

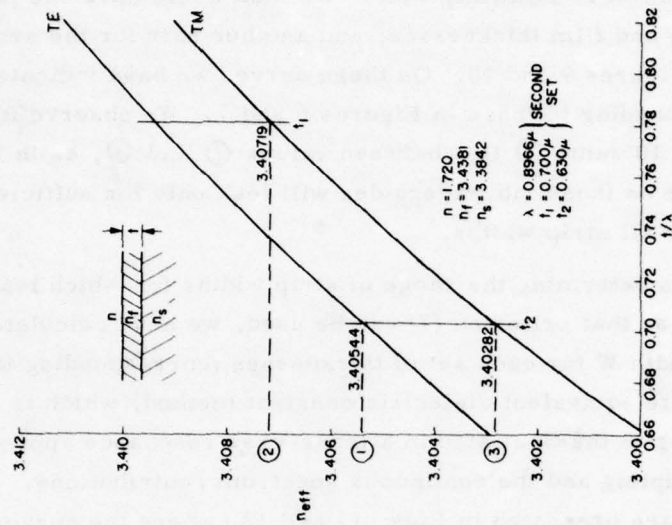


Fig. 9. Dispersion curves of the type shown in Fig. 4 corresponding to the strip and film regions of a set of specific rib waveguides. The pertinent numerical parameters are shown in the figure.

Fig. 10. Dispersion curves similar to those of Fig. 9 for a second set of rib waveguides.

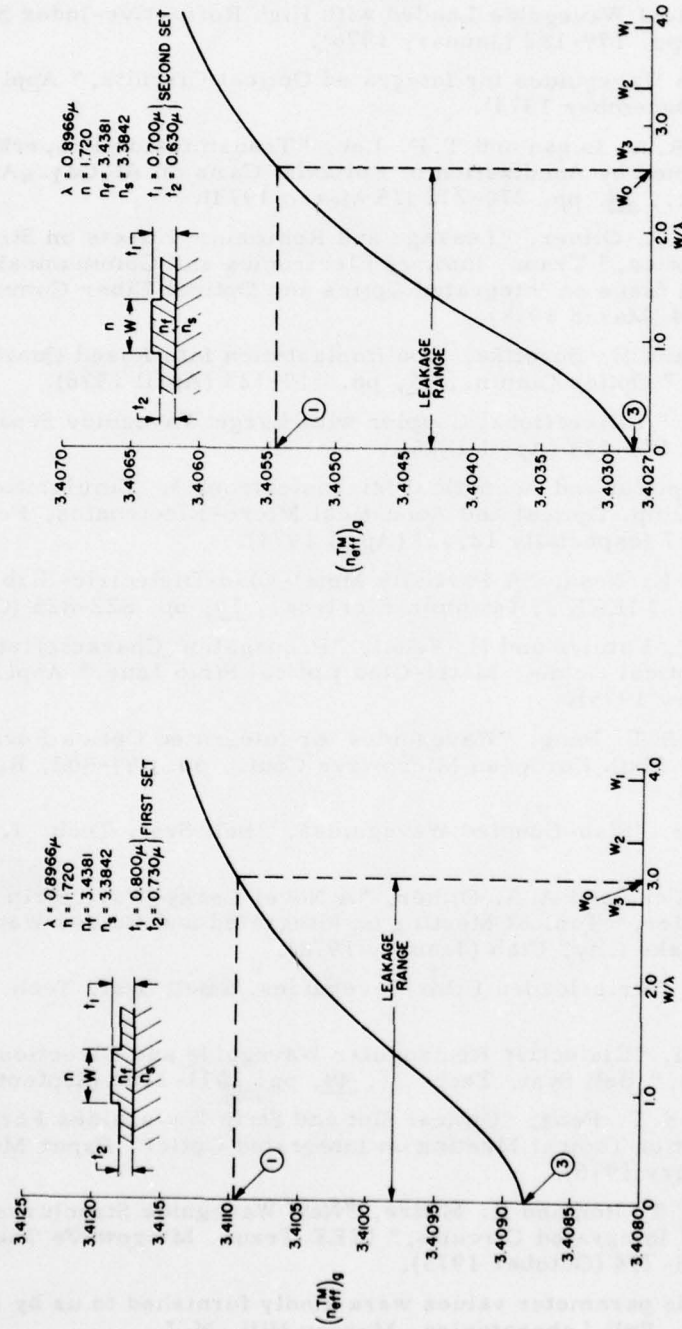


Fig. 11. Dispersion plot, of the type shown in Fig. 8, for a set of specific rib waveguides. The strip widths  $W_1$ ,  $W_2$  and  $W_3$  are indicated on the horizontal axis.

Fig. 12. Dispersion plot similar to that in Fig. 11 for a second set of rib waveguides. The strip widths  $W_1$ ,  $W_2$  and  $W_3$  are indicated on the horizontal axis.

## REFERENCES

1. H. Furuta, H. Noda and A. Ihaya, "Novel Optical Waveguide for Integrated Optics," *Appl. Opt.*, 13, pp. 322-326 (February 1974).
2. N. Uchida, "Optical Waveguide Loaded with High Refractive-Index Strip Film," *Appl. Opt.*, 15, pp. 179-182 (January 1976).
3. J.E. Goell, "Rib Waveguides for Integrated Optical Circuits," *Appl. Opt.*, 12, pp. 2797-2798 (December 1973).
4. F.K. Reinhart, R.A. Logan and T.P. Lee, "Transmission Properties of Rib Waveguides Formed by Anodization of Epitaxial GaAs on  $\text{Al}_x\text{Ga}_{1-x}\text{As}$  Layers," *Appl. Phys. Lett.*, 24, pp. 270-272 (15 March 1974).
5. S.T. Peng and A.A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics," *Trans. Inst. of Electronics and Communication Engineers of Japan, Special Issue on Integrated Optics and Optical Fiber Communications*, E61, pp. 151-154 (March 1978).
6. R. Th. Kersten and H. Boroffka, "Ion Implantation Into Fused Quartz for Integrated Optical Circuits," *Optics Comm.*, 17, pp. 119-123 (April 1976).
7. R. Th. Kersten, "A Directional Coupler with Large Waveguide Separation," *Optics Comm.*, 17, pp. 124-128 (April 1976).
8. A.A. Oliner, "Optical and Acoustical Microelectronics: Similarities and Differences," *Proc. Symp. Optical and Acoustical Micro-Electronics*, Polytechnic Inst. of N.Y., pp. 1-17 (especially 12, 13) (April 1974).
9. J. Hamasaki and K. Nosu, "A Partially Metal-Clad-Dielectric-Slab Waveguide for Integrated Optics," *IEEE J. Quantum Electron.*, 10, pp. 822-825 (October 1974).
10. Y. Yamamoto, T. Kamiya and H. Yanai, "Propagation Characteristics of a Partially Metal-Clad Optical Guide: Metal-Clad Optical Strip Line," *Appl. Opt.*, 14, pp. 322-326 (February 1975).
11. A.A. Oliner and S.T. Peng, "Waveguides for Integrated Optics Formed by Metal Platings," *Proc. Sixth European Microwave Conf.*, pp. 499-503, Rome, Italy (September 1976).
12. E.A.J. Marcatili, "Slab-Coupled Waveguides," *Bell Syst. Tech. J.*, 53, pp. 645-674 (April 1974).
13. E.W. Hu, S.T. Peng and A.A. Oliner, "A Novel Leaky-Wave Strip Waveguide Directional Coupler," *Topical Meeting on Integrated and Guided Wave Optics*, Paper No. WD2, Salt Lake City, Utah (January 1978).
14. V. Ramaswamy, "Strip-loaded Film Waveguides," *Bell Syst. Tech. J.*, 53, p. 697 (April 1974).
15. E.A.J. Marcatili, "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," *Bell Syst. Tech. J.*, 48, pp. 2071-2102 (September 1969).
16. A.A. Oliner and S.T. Peng, "Optical Slot and Strip Waveguides Formed by Metal Platings," *Digest of Topical Meeting on Integrated Optics*, Paper MC5, Salt Lake City, Utah (January 1976).
17. W.V. McLevige, T. Itoh and R. Mittra, "New Waveguide Structures for Millimeter Wave and Optical Integrated Circuits," *IEEE Trans. Microwave Theory Tech.*, MTT-23, pp. 788-794 (October 1975).
18. The rib waveguide parameter values were kindly furnished to us by F.K. Reinhart and J.C. Shelton, Bell Laboratories, Murray Hill, N.J.

## MODULATION SENSITIVITY OF ELECTRO-OPTIC SLAB WAVEGUIDES

S. T. Peng

The advantage of economy in power and size afforded by thin-film waveguiding structures for electro-optic and acousto-optic modulation and beam deflection has been well recognized.<sup>1,2,3</sup> The central question concerning such devices is: what is the incremental phase change of a confined mode under the action of an applied electric or acoustic field? Such a question was first explored by Lotspeich for a single layer structure<sup>1</sup> and the results were extended by Buckman for multilayer structures.<sup>2,3</sup> These previous analyses were primarily based on a numerical analysis which is time consuming and, more importantly, lacks physical insight into the wave phenomenon associated with the devices. So far, no systematic design criterion has been put forth in the literature. In this work, we have adopted a new approach and successfully derived simple analytic formulas for the modulation sensitivity of a structure with any number of layers. As an illustration, the results for a TE mode in a single layer structure are given in what follows.

A thin film electro-optic modulator is shown schematically in Figure 1. The electro-optic material of refractive index  $n_f$  and thickness  $t$  is sandwiched between

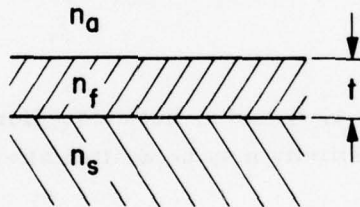


Fig. 1. Thin film electro-optic modulator.

two uniform media of refractive indices  $n_a$  and  $n_s$ , respectively. Without loss of generality, the refractive indices may be normalized such that  $n_a = 1$ , and the space above the film may then be referred to as the air region and the space below as the substrate region. Therefore, the subscripts a, f, and s designate the air, film and substrate regions, respectively. For the confinement of optical energy within the thin film, the refractive indices must satisfy  $n_f > n_s > n_a$ .

It is well known that the propagation characteristic of surface waves along the thin film are completely determined by a dispersion relation (or eigen equation) of the electromagnetic boundary value problem. It can be shown that the dispersion relation may be cast into many different mathematical forms; with physical insight into the basic surface-wave phenomenon we have observed that one of them is particularly useful for the study of electro-optic modulators. Such an observation leads

to the derivation of simple analytic formulas by which the analysis and design of electro-optic modulators can be easily performed.

The variable of interest here is the modulation sensitivity of the active thin film waveguide structure:<sup>1,2</sup>  $S = d\beta_g/dk_f$ , where  $\beta_g$  is the propagation constant of a guided surface-wave mode and  $k_f$  is the plane wave propagation constant of the electro-optic medium. If we define the effective index of refraction:

$$n_g = \beta_g/k_0 \quad (1)$$

where  $k_0$  is the plane wave propagation constant in free space, we then have the modulation sensitivity:

$$S = \frac{d\beta_g}{dk_f} = \frac{dn_g}{dn_f} \quad (2)$$

for which we have made use of the relations:

$$k_r = k_0 n_r, \quad \text{for } r = a, f, \text{ and } s. \quad (3)$$

We observe that instead of the indices of refraction, it will be more convenient to analyze the electro-optic modulators in terms of the dielectric constants:

$$\epsilon_r = n_r^2 \quad (4)$$

where  $r$  may stand for not only  $a$ ,  $f$ , and  $s$ , but also  $g$  (for guided modes). From Eqs. (2) and (4), the modulation sensitivity may be written alternatively in the form:

$$S = \frac{dn_g}{dn_f} = \frac{n_f}{n_g} \frac{d\epsilon_g}{d\epsilon_f}. \quad (5)$$

For the structure shown in Fig. 1, it is well known that  $n_s \leq n_g < n_f$ . Consequently, we have:

$$1 < \frac{n_f}{n_g} \leq \frac{n_f}{n_s}. \quad (6)$$

Therefore, the first factor of the modulation sensitivity given by Eq. (5) is always greater than unity and is bounded above by the ratio  $n_f/n_s$ . We have derived simple analytic expressions for the other factor,  $d\epsilon_g/d\epsilon_f$ , for both TE and TM polarizations. For example, in the case of an active layer in free space ( $n_s = n_a$ ), we have, for the TE polarization:

$$\frac{d\epsilon_g}{d\epsilon_f} = 1 - f(\epsilon_g) \quad (7)$$

with

$$f(\epsilon_g) = \frac{\cos^2[k_o t \sqrt{\epsilon_f - \epsilon_g}]}{1 + k_o t \sqrt{\epsilon_g - \epsilon_a}} \quad (8)$$

where  $f$  as a function of  $\epsilon_g$  is obviously bounded below and above by:

$$0 < f(\epsilon_g) \leq 1, \quad (9)$$

for any guided TE mode of the structure under consideration. Evidently, it then follows from Eqs. (7) and (9) that

$$0 \leq \frac{d\epsilon_g}{d\epsilon_f} < 1. \quad (10)$$

It is noted that the equality sign in Eqs. (5), (8), and (9), holds if and only if a mode is right at the cutoff condition ( $n_g = n_a$ ). Although analytic expressions for the factor  $d\epsilon_g/d\epsilon_f$  become more complicated if the substrate is present, the inequalities in Eq. (10) hold in any case. In view of Eqs. (6) and (10), it is evident from Eq. (5) that we have:

$$0 \leq S < n_f/n_s. \quad (11)$$

Again, the equality sign holds only at the cutoff condition. Therefore, we can now draw the conclusion: for a single layer of electro-optic material on a substrate, the modulation sensitivity can never be equal to or exceed the ration  $n_f/n_s$ .

It is noted that, in contrast to the published approach<sup>1-3</sup> based on a dispersion relation in terms of phase functions, our analysis is based on one in terms of impedance function. The advantage of this different approach is that it yields simple analytic expressions for the modulation sensitivity for multilayer structure; this opens the way for an optimized synthesis procedure for thin-film modulators.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

S. T. Peng

#### REFERENCES

1. J. F. Lotspeich, "Modulation-Sensitivity Enhancement in Electro-Optic Slab Waveguides," J. Optic. Soc. Am., Vol. 65, pp. 797-803 (July 1975).
2. A. B. Buckman, "Effective Electro-Optic Coefficient of Multilayer Dielectric Waveguides: Modulation Enhancement," J. Optic. Soc. Am., Vol. 66, pp. 30-34 (January 1976).
3. A. B. Buckman, "Theory of an Efficient Electronic Phase Shifter Employing a Multilayer Dielectric-Waveguide Structure, IEEE Trans. MTT, Vol. MTT-25, pp. 480-483 (June 1977).

## RAY ANALYSIS OF UNSTABLE RESONATORS

S. H. Cho and L. B. Felsen

The ray analysis of high-frequency propagation and scattering problems is successful especially when the required number of constituent ray fields is not very large. This is often the case when an impenetrable scatterer is in a free space environment since multiple diffraction phenomena between scattering centers, if these exist, decrease rapidly in intensity. However, when the propagation environment includes multiple boundaries, between which fields can be reflected an arbitrarily large number of times without substantial loss, the resulting "ray series" becomes cumbersome and inaccurate. Therefore, the success of the ray method in the presence of multiple reflectors depends on one's ability to treat all higher order reflected rays collectively so that the ray series can be truncated at a tractable limit. If the reflector surfaces are infinite in extent, thereby forming a waveguide or duct, the propagating fields can alternatively be expressed in terms of modes guided, loosely speaking, in a direction parallel to the boundaries. Guided modes can be represented in terms of modal ray congruences (families of rays)<sup>1</sup> and the associated wavefronts (equiphase surfaces), one upgoing and one downgoing, such that the boundary conditions on the reflector surfaces are satisfied (Fig. 1). To represent a modal field, these ray congruences

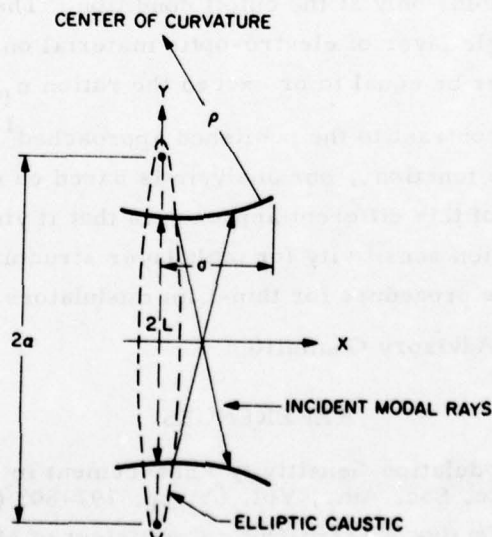


Fig. 1. Ray congruences for self-consistent fields between infinite hyperbolic mirrors.

must be self-consistent in the sense that the upgoing congruence, on reflection at the upper boundary, is converted into the downgoing congruence which, after reflection at the lower boundary, again becomes the original upgoing congruence. Between curved boundaries, these congruences are generated by "modal caustics" whose location is such as to meet the self-consistency constraint. The self-consistent caustics furnish, equivalently, the eigenvalues for the guided modes. A special example of these modal congruences and wavefronts in the unstable resonator environment is furnished by Siegman's<sup>2</sup> two spherical waves emanating from the foci of an elliptic coordinate system when the boundaries are infinite hyperboloidal mirrors; since the mirrors are actually truncated near the apex, they can be approximated there by spherical shapes.

The ability to represent fields between infinite mirrors by two modal ray congruences suggests that source or diffraction excited ray fields with many reflections may be combined collectively into modal congruence form. Our confirmation of this conjecture has rendered the ray-optical treatment of the unstable resonator feasible. The analysis proceeds by assuming an incident ray field (modal congruence) appropriate to the infinite mirror structure. When the mirrors are truncated, this incident field excites the mirror edges which thereby become scattering centers for GTD (geometrical theory of diffraction) edge diffracted rays. These rays propagate back into the resonator via direct (Fig. 2a) and multiply reflected paths (Fig. 2b), with amplitudes and phases determined according to the rules of GTD; they modify the incident field on the mirrors.

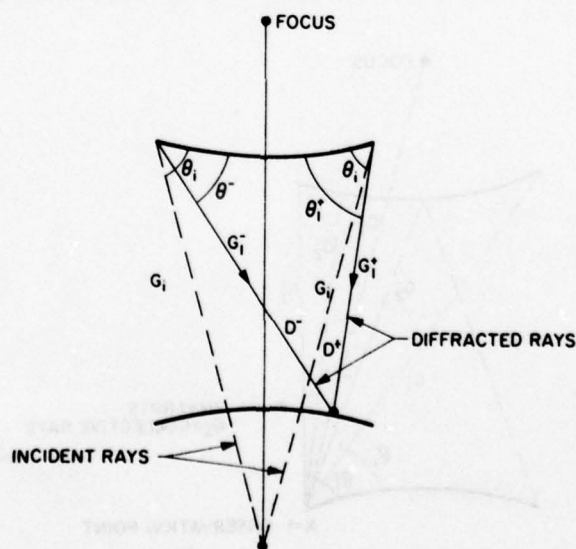


Fig. 2a. Diffraction of incident modal field.

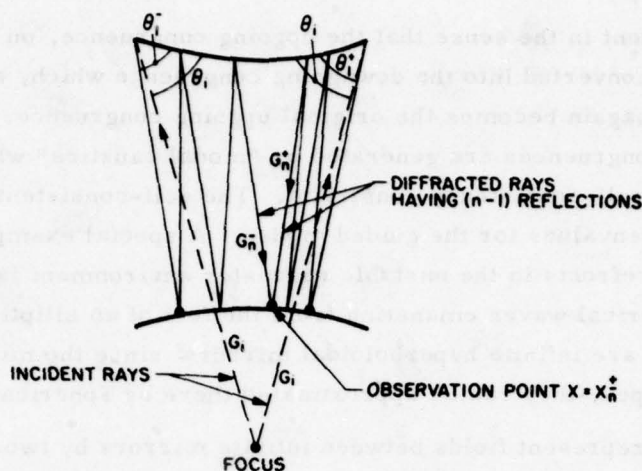


Fig. 2b. Multiple reflection of primary diffracted field

Although a ray can undergo an arbitrarily large number of reflections between the mirrors before reaching an observation point, the collective behavior of all high-order reflected rays with more than  $N$  reflections reduces to  $(2N+1)$  the number of rays reaching an observation point. The choice of  $N$  is not critical but must be large enough to meet the "collective (i.e., modal) ray" criterion.

When, after traversing the resonator, each of the multiply reflected edge diffracted rays impinges on the edges along its geometrically determined incidence angle (Fig. 2c), each ray excites  $(2N+1)$  secondary edge-diffracted rays. These rays propagate through the resonator like the primary diffracted rays and provide secondary

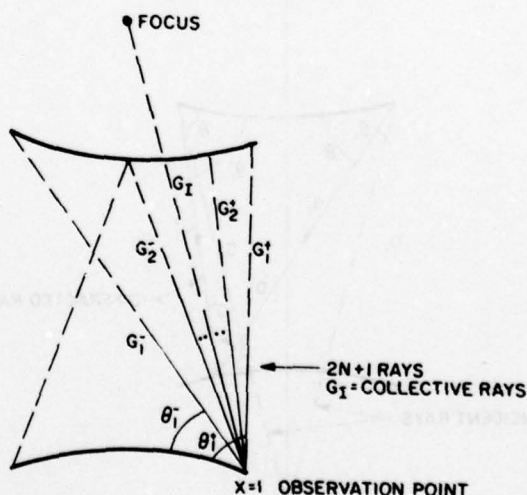


Fig. 2c. Total field at edge after primary diffraction.

incident fields at the edges, along  $(2N+1)$  trajectories whose directions coincide with those of the previous diffraction order. The same comment applies to any higher order of diffraction. At this stage of the analysis, one may seek resonant field solutions by demanding that the edge excitation for two successive orders of diffraction

are the same (Fig. 3). To satisfy this self-consistency requirement, which leads to a resonance equation, it is necessary to adjust the incidence angles of the  $(2N+1)$  rays at the edges, including that of the collective modal ray. This, in turn, implies the proper selection of the modal caustic parameter, which, in fact, becomes the eigenvalue of the resonant mode. After the eigenvalues have been found from solution of the resonance equation, the corresponding resonant modal field is obtained as the sum of the  $(2N+1)$  multiply reflected ray fields.

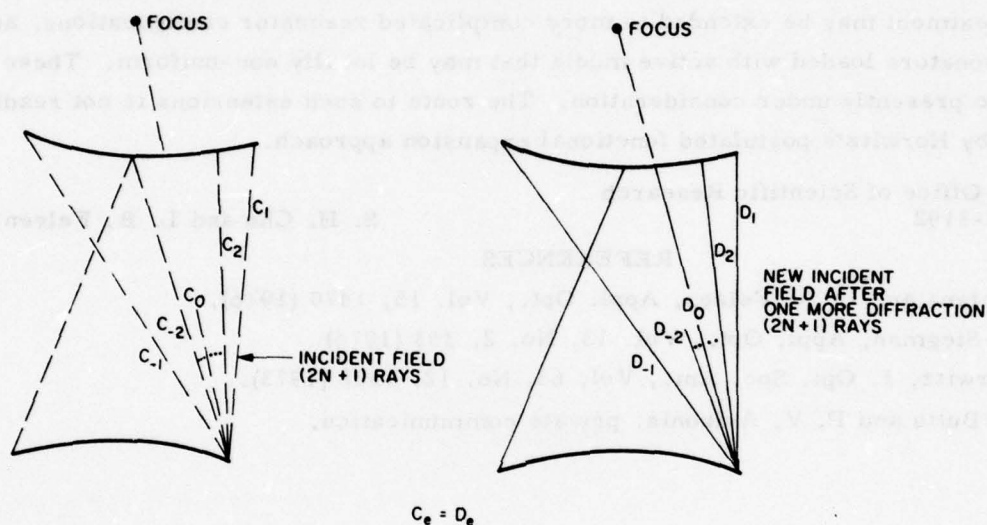


Fig. 3. Resonance condition.

We have applied the above procedure to symmetrical strip and circular mirror resonators. After making an approximation, our ray-optically derived resonance condition can be simplified to agree completely with that derived by Horwitz<sup>3</sup> from a postulative approach based on an asymptotic treatment of the resonator integral equation. Since the approximations made at various stages of the ray-optical analysis have been stated carefully, it is possible to identify what must be done if the present results are to be improved. Moreover, the direct comparison of quantities and mathematical relations appearing in our study with those of Horwitz permits the deductive understanding of Horwitz's postulative approach and provides a physical basis for his mathematically convenient assumptions at various stages. A comparison has also been made with the resonance equation obtained by the waveguide technique of Santana and Felsen,<sup>1</sup> and it has been shown that approximate analytical solutions obtained by these two different procedures for the lowest mode eigenvalues at adequately large Fresnel numbers are in agreement. By combining a collective with an individual treatment of ray contributions, the present method thus furnishes a link between the ray-optical and waveguide schemes.

The complex eigenvalues and the corresponding eigenmode fields calculated from equations derived by the Horwitz method have been found to agree well with those generated by direct numerical solution of the resonator integral equation. The importance of the present study is not limited to providing a deductive derivation of these earlier results for idealized two-dimensional and three-dimensional shapes. Because of the generality of the ray method, which can account for imperfect mirror geometries and for propagation phenomena ascribed to local medium inhomogeneities, it is suggestive that our treatment may be extended to more complicated resonator configurations, and also to resonators loaded with active media that may be locally non-uniform. These aspects are presently under consideration. The route to such extensions is not readily perceived by Horwitz's postulated functional expansion approach.

Air Force Office of Scientific Research  
AFOSR-77-3192

S. H. Cho and L. B. Felsen

#### REFERENCES

1. C. Santana and L. B. Felsen, Appl. Opt., Vol. 15, 1470 (1976).
2. A. E. Siegman, Appl. Opt., Vol. 13, No. 2, 353 (1974)
3. P. Horwitz, J. Opt. Soc. Am., Vol. 63, No. 12, 1528 (1973).
4. R. R. Butts and P. V. Avizonis, private communication.

## SMALL MISALIGNMENT EFFECTS IN UNSTABLE RESONATORS

C. Santana and L. B. Felsen

A recently published communication<sup>1</sup> contains a numerical study of the integral equation for unstable strip resonators with misaligned sharp-edged circular mirrors. The results, which cover only a limited range of Fresnel numbers, indicate that even very small misalignments can have a marked effect on the resonant mode loss behavior. In particular, the misalignment leads to mode detachment at low Fresnel numbers and produces additional oscillations in the eigenvalue curve. The regularity of these oscillations led the authors<sup>1</sup> to suggest that this behavior would be valid also for other parameter values, and to indicate a criterion for choosing Fresnel numbers that minimize the misalignment effects. We show here that when the eigenvalues are examined over a broader range of Fresnel numbers, the conclusions based on the data in Ref. 1 are misleading. As seen from Fig. 1, which contains our data (solid curve) and those from

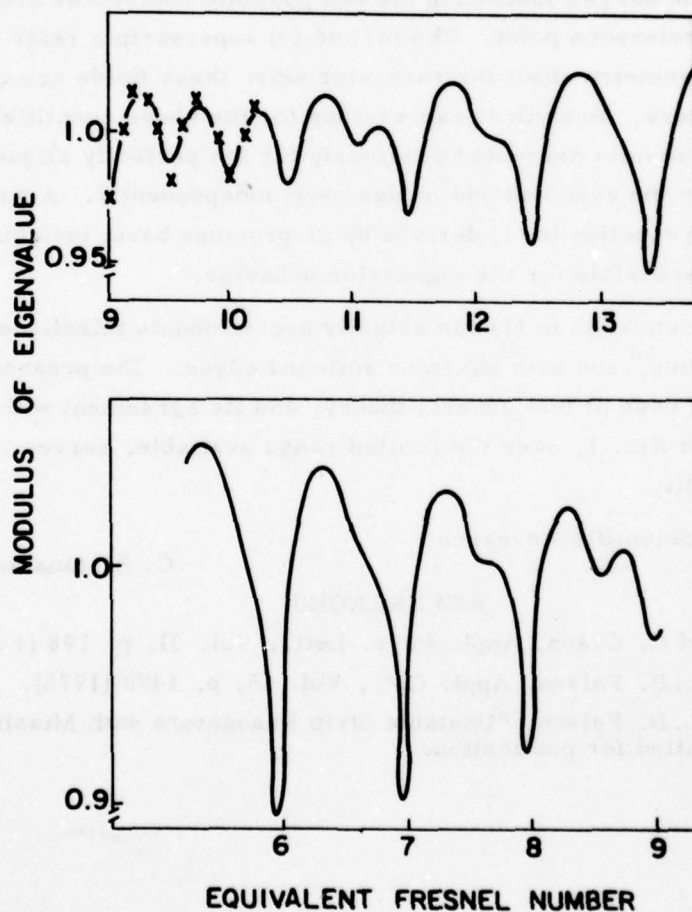


Fig. 1. Eigenvalue behavior for lowest order detached mode. Misalignment parameter  $\epsilon = 0.01262$  (see Ref. 1). The eigenvalues are normalized to the geometric optical value for infinite mirrors.

Ref. 1 (shown as crosses), the eigenvalue plot loses its regularity and exhibits changed periodicities. These features can be explained physically by the alternative method, summarized below, that was used in our analysis. We have also calculated, and shall interpret, data for larger misalignments than those given in Reference 1.

The solid curve in Fig. 1 was produced by solution of the resonance equation

$$[\bar{R}_n^{(e)} \bar{\Gamma}_{nn} - 1] [\bar{R}_n^{(o)} \bar{\Gamma}_{nn} - 1] + [\bar{R}_n^{(o)} \bar{\Gamma}_{nn} - 1] [\bar{R}_n^{(e)} \bar{\Gamma}_{nn} - 1] = 0 \quad (1)$$

which is obtained when the resonator is viewed as a waveguide in the direction transverse to its axis.<sup>2</sup> In deriving Eq. (1), the tilted mirror geometry is first mapped into a configuration with non-tilted mirrors, which are asymmetrical with respect to the resonator axis. In this configuration,  $\bar{R}_n$  and  $\bar{\Gamma}_{nn}$  identify the modal reflection coefficients<sup>2</sup> due to the modal caustic at the center of the resonator and the mirror edges, respectively, with the arrows indicating the two possible transverse directions seen from a preselected reference point. The (e) and (o) superscripts refer to modal fields with even and odd symmetry about the resonator axis; these fields are coupled in the asymmetrical structure. Analytical expressions for the above quantities are derived by generalization of results presented previously for the perfectly aligned symmetrical configuration,<sup>2</sup> where the even and odd modes exist independently. A simplified version of the full resonance equation in (1) derived by us provides basic insights into the physical mechanisms responsible for the eigenvalue behavior.

The resonance equation in (1) can actually accommodate misaligned mirrors of unequal size and radius,<sup>3</sup> and with sharp or softened edges. The present example constitutes a special case of this general theory, and its agreement with the independently obtained results in Ref. 1, over the limited range available, serves as further confirmation of its validity.

Air Force Office of Scientific Research  
AFOSR-77-3192A

C. Santana and L. B. Felsen

#### REFERENCES

1. J. F. Perkins and C. Cason, Appl. Phys. Lett., Vol. 31, p. 198 (1 August 1977).
2. C. Santana and L. B. Felsen, Appl. Opt., Vol. 15, p. 1470 (1976).
3. C. Santana and L. B. Felsen, "Unstable Strip Resonators with Misaligned Circular Mirrors," submitted for publication.

## PROPAGATION IN INHOMOGENEOUS SLAB WAVEGUIDES -- EXACT SOLUTIONS

E. Navon and L. B. Felsen

A. Introduction

Slab waveguides with an inhomogeneous dielectric profile embedded in a homogeneous environment find application in integrated optics, and they also serve as prototype models for graded index optical fibers. Solutions for the modal fields and modal propagation coefficients in these configurations must generally be determined by approximate methods suitable for the high-frequency range. In order to check the validity of these methods and the accuracy of the results, it is important to have exact solutions for special profiles, available for comparison. Here, we examine two such special profiles, the parabolic and hyperbolic tangent. The formulation and analysis is performed in such a manner that the procedure becomes applicable also to more general profiles, which are to be treated by the recently developed evanescent wave tracking method. This method has been shown to yield high frequency asymptotic modal solutions when the inhomogeneous refractive index profile extends to infinity.<sup>1</sup> By comparison with the exact solutions developed below, the evanescent wave tracking method is being extended to account for truncated profiles within a homogeneous surround.

The truncated parabolic profile has been treated previously, but by a procedure less suited to our purpose than the one presented here.<sup>2</sup> Our solutions for the truncated hyperbolic tangent profile are new.

B. General Formulation

We seek solutions  $u(x, z)$  of the scalar wave equation

$$[\nabla^2 + k^2 n^2(x)] u(x, z) = 0, \quad \partial/\partial y \equiv 0 \quad (1)$$

where  $k$  is the free space wavenumber and  $n(x)$  is a real refractive index. An  $\exp(-i\omega t)$  time dependence is suppressed. These source-free (modal) fields should be bounded in the entire interval  $-\infty < x < \infty$ , and therefore decay for large  $x$ , and they propagate along the  $z$ -direction. The guiding medium consists of an inhomogeneous but symmetric inner (core) region wherein  $n(x)$  behaves parabolically near the  $x=0$  plane; the core is surrounded by an infinite homogeneous exterior medium (see Figure 1). At the boundaries  $|x| = x_c$  between these regions, the refractive index may have a finite discontinuity. Thus  $n^2(x)$  is given by

$$n^2(x) = \begin{cases} n_0^2 - \bar{n}^2(x) & |x| \leq x_c \\ n_2^2 & |x| > x_c \end{cases} \quad (2)$$

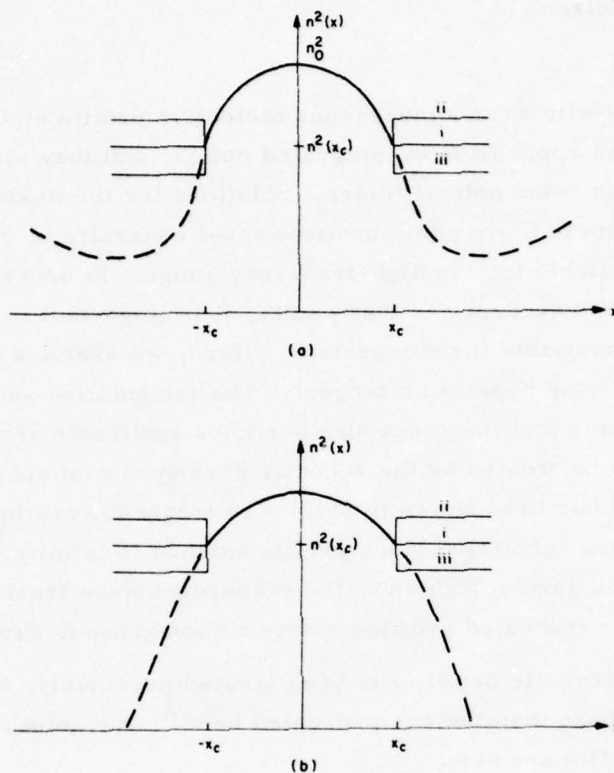


Fig. 1. Various profile shapes. (a) General symmetrical profile; (b) parabolic profile. --- unbounded profile, — bounded profile with (i) continuous  $n_2 = n(x_c)$ ; (ii)  $n_2 > n(x_c)$ ; (iii)  $n_2 < n(x_c)$ .

where  $n_0 \equiv n(0)$ , and  $\bar{n}(x)$  is a function description of the profile. For the special case of a continuous profile, one has  $n_2 = n(x_c)$ .

We shall consider first the solution for an infinite core wherein  $x_c \rightarrow \infty$ . We seek to determine the discrete set of modal propagation coefficients  $\beta_l$ , which will be found to depend on the integer  $l$  that also serves to distinguish various modes. When  $x_c$  is finite, the presence of the exterior region modifies the propagation coefficients in that the integer  $l$  for the unbounded core is replaced by a non-integral index  $\nu$ :

$$\nu = l + \Delta\nu_l \quad (3)$$

Accordingly, the propagation coefficients  $\beta_l$  undergo the transformation  $\beta_l \rightarrow \beta_\nu$ . The dispersion equation for  $\Delta\nu_l$  is derived by imposing the boundary conditions at  $|x| = x_c$ .

We shall also require a consistent procedure for uniquely normalizing the eigenfunctions. The construction of even and odd solutions, as employed by Hashimoto,<sup>2</sup> is found to be helpful. However, we define these solutions in terms of known functions, which have series expansions valid in all regions of  $x$ , rather than in terms of the integral representations used previously.<sup>2</sup> The series expansion forms are found to have advantages of analytic simplicity and computational convenience.

In accord with the preceding remarks, the wave function  $u(x, z)$ , to be identified with the transverse electric field  $E_y$ , is expressed as

$$u_v(x, z) = f_v(x) \exp(i\beta_v z) \quad (4)$$

where  $\beta_v$  is the propagation coefficient in the  $z$  direction. Substitution into the wave equation (1) yields a second order ordinary differential equation for  $f_v(x)$ :

$$\left[ \frac{d^2}{dx^2} + k^2 n^2(x) - \beta_v^2 \right] f_v(x) = 0 \quad (5)$$

which has the general solution

$$f_v(x) = A_v \phi_v(x) + D_v \psi_v(x) \quad (6)$$

where  $A_v$  and  $D_v$  are constants independent of  $x$ , and  $\phi_v(x)$  and  $\psi_v(x)$  are linearly independent solutions of the differential equation (5), which shall be chosen as decaying and growing solutions, respectively, as  $|x| \rightarrow \infty$ .

### 1. Unbounded Core

The refractive index profile for the unbounded medium is of the form

$$n^2(x) = n_0^2 - \bar{n}^2(x) \quad -\infty < x < \infty \quad (7)$$

For guided modes, which decay at  $|x| \rightarrow \infty$ , it is necessary to eliminate the growth function  $\psi_v(x)$ , whence  $D_v \equiv 0$  in Equation (6). Furthermore, it is assumed that the function  $\phi_v(x)$  is now characterized by integral values of the parameter  $v$  so that  $v = \ell = 0, 1, 2, \dots$ . Thus, the modal field solutions for the unbounded profile becomes

$$f_v(x) = f_\ell(x) = A_\ell \phi_\ell(x) \quad (8)$$

with discrete eigenvalues  $\beta_\ell$ .

## 2. Bounded Core

For the bounded core profile in Eq. (2), the assumed solution is of the form

$$u_v(x, z) = \begin{cases} f_v(x) \exp(i\beta_v z) & |x| \leq x_c \\ f_v(x_c) \exp(-\alpha_v |x - x_c|) \exp(i\beta_v z) & |x| > x_c \end{cases} \quad (9)$$

where  $\alpha_v = [\frac{\partial}{\partial x} u_v(x, z)/u_v(x, z)]_{x=x_c^+}$  is the transverse wavenumber in the exterior region,

$$\alpha_v = [\beta_v^2 - n^2(x_c^+)]^{1/2} \quad (10)$$

and  $x_c^+$  is on the exterior side of the boundary ( $x_c^+ = x_c + \delta$ ,  $\delta > 0$ ,  $\delta \rightarrow 0$ ). A choice of positive  $\alpha_v$  assures decay as  $|x| \rightarrow \infty$ .

The solution in the core region is the same as in Equation (6). However, now  $D_v \neq 0$ , and  $v$  is not required to be an integer but is determined from the boundary conditions at  $|x| = x_c$ : continuity of the tangential components of electric field ( $u_v$ ) and magnetic field ( $\propto \partial u_v / \partial x$ ). Continuity of  $u_v$  is already satisfied by the assumed form in Equation (9). Continuity of  $\partial u_v / \partial x$  leads to the dispersion equation

$$f_v'(x_c) + \alpha_v f_v(x_c) = 0 \quad (11)$$

where the prime denotes differentiation with respect to  $x$ . Here,  $f_v(x)$  implies that in the expression for  $\phi_\ell(x)$ ,  $\psi_\ell(x)$  and  $\beta_\ell$  for the unbounded core, the integer  $\ell$  is replaced by  $v$ . The previously discarded growing solution  $\psi_\ell(x)$  must now be included in the analysis.

To simplify solution of the dispersion equation, and to exhibit the effect of truncation of the core medium as a perturbation about the unbounded case, the normalization coefficients  $A_v$  and  $D_v$  in Eq. (6) should be chosen appropriately. Since the refractive index profile considered here is symmetric with respect to  $x$ , the modal fields separate into two sets having even and odd  $x$ -symmetry, respectively. Following Hashimoto,<sup>2</sup> we shall normalize the functions  $\phi_v(x)$  and  $\psi_v(x)$  in such a manner that  $f_v(x)$  can be expressed as

$$f_v^o(x) = \begin{Bmatrix} \cos \frac{\pi v}{2} \\ \sin \frac{\pi v}{2} \end{Bmatrix} \phi_v(x) + \begin{Bmatrix} -\sin \frac{\pi v}{2} \\ \cos \frac{\pi v}{2} \end{Bmatrix} \psi_v(x) \quad (12)$$

where superscripts  $e$  and  $o$  designate even and odd modes, respectively. When  $v$  takes on even or odd integer values  $\ell$ , this solution reduces to that in Eq. (8) for the unbounded profile. Explicit retention of the decaying function  $\phi_v$  and the growing function  $\psi_v$  in

the even or odd modal field solutions is useful not only for describing the perturbing effect of the exterior medium but also for the subsequent development of an asymptotic theory for determination of  $f_v$ . For the special profiles discussed below, the known even and odd solutions of Eq. (5) can be manipulated into the canonical form Eq. (12) by use of relevant circuital relations, as will be demonstrated. For the parabolic profile, Hashimoto<sup>2</sup> obtained the split-up in Eq. (12) by a more complicated manipulation of contour integral representations.

Using Eq. (12), one obtains from Eqs. (11) and (3) for mode fields with even or odd x-symmetry, for any integral  $l$ :

$$\tan \frac{\pi \Delta v_l}{2} = \frac{\phi'_v(x_c) + \alpha_v \phi_v(x_c)}{\psi'_v(x_c) + \alpha_v \psi_v(x_c)} \quad (13)$$

For modes with  $\Delta v_l \ll 1$ , this transcendental equation can be solved approximately by assuming that  $\phi_v(x) \approx \phi_l(x)$  and  $\psi_v(x) \approx \psi_l(x)$ , and by replacing the tangent by its argument:

$$\Delta v_l \approx \frac{2}{\pi} \frac{\phi'_l(x_c) + \alpha_l \phi_l(x_c)}{\psi'_l(x_c) + \alpha_l \psi_l(x_c)} \quad (14)$$

This approximation is generally applicable to lower order modes which are minimally affected by the exterior medium. It also can be employed as the first iteration in a numerical scheme for determining improved values of  $\Delta v_l$  from Equation (13).

### C. Parabolic Profile

The two dimensional parabolic profile is defined as (see Fig. 1)

$$n^2(x) = \begin{cases} n_0^2 - a_0^2 x^2 & |x| \leq x_c \\ n_2^2 & |x| > x_c \end{cases} \quad (15)$$

where  $a_0$  and  $n_2$  are constants and  $n_0 \equiv n(0)$ . Thus, Eq. (6) becomes for the uncladded medium ( $x_c \rightarrow \infty$ ):

$$\left[ \frac{d^2}{d\eta^2} - 2\eta \frac{d}{d\eta} + 2l \right] g_l(\eta) = 0 \quad (16)$$

where

$$g_l(\eta) \equiv f_l(\eta) \exp(\eta^2/2) \quad (17)$$

$$\eta \equiv \sqrt{ka_0} x, \quad \eta_c = \sqrt{ka_0} x_c \quad (18)$$

with the solution (bounded at  $|x| \rightarrow \infty$ )<sup>3</sup>

$$g_\ell(\eta) = H_\ell(\eta) \quad (19)$$

$$\beta_\ell = n_0 k \left[ 1 - \frac{a_0(2\ell + 1)}{kn_0^2} \right]^{1/2} \quad (20)$$

Here,  $H_\ell(\eta)$  is the Hermite polynomial of the  $\ell$ -th order, and  $\ell = 0, 1, 2, \dots$  is an integer that can be interpreted as the guided mode index.

To deal with the cladded profile Eq. (15), we change  $\ell$  into  $\nu = \ell + \Delta\nu_\ell$  in Equation (20). The modified equation (16) now has the two linearly independent solutions  $g_\nu(\eta) = H_\nu(\eta)$  and  $h_\nu(\eta) = h_\nu(\eta) \equiv \exp(\eta^2) H_{-\nu-1}(i\eta)$ , which are the Hermite functions of the first and second kind, respectively.<sup>3</sup> These functions can be represented as follows:<sup>4,†</sup>

$$H_\nu(\eta) = \pi^{-1/2} \left[ \cos \frac{\pi\nu}{2} \Gamma\left(\frac{1}{2} + \frac{\nu}{2}\right) {}_1F_1\left(-\frac{\nu}{2}, \frac{1}{2}, \eta^2\right) + 2\eta \sin \frac{\pi\nu}{2} \Gamma\left(1 + \frac{\nu}{2}\right) {}_1F_1\left(\frac{1}{2} - \frac{\nu}{2}, \frac{3}{2}, \eta^2\right) \right] \quad (21a)$$

$$h_\nu(\eta) = \pi^{-1/2} \left[ -\sin \frac{\pi\nu}{2} \Gamma\left(\frac{1}{2} + \frac{\nu}{2}\right) {}_1F_1\left(-\frac{\nu}{2}, \frac{1}{2}, \eta^2\right) + 2\eta \cos \frac{\pi\nu}{2} \Gamma\left(1 + \frac{\nu}{2}\right) {}_1F_1\left(\frac{1}{2} - \frac{\nu}{2}, \frac{3}{2}, \eta^2\right) \right] \quad (21b)$$

where  ${}_1F_1$  is the confluent hypergeometric function defined as

$${}_1F_1(a, c, z) \equiv \sum_{m=0}^{\infty} \frac{(a)_m}{m! (c)_m} z^m; \quad (a)_m \equiv a(a+1) \cdots (a+m-1) = \frac{\Gamma(a+m)}{\Gamma(a)}$$

and  $\Gamma(q)$  is the gamma function with argument  $q$ . The Wronskian of the Hermite function is<sup>4†</sup>

$$W_\eta [H_\nu(\eta), h_\nu(\eta)] = \pi^{-1/2} 2^{1-\nu} \Gamma(1+\nu) \exp(\eta^2) \quad (22)$$

One may now define  $g_\nu^{(e)}(\eta)$  and  $g_\nu^{(o)}(\eta)$  as even and odd functions with respect to  $\eta$ :

$$g_\nu^{(e)}(\eta) \equiv \pi^{-1/2} \Gamma\left(\frac{1}{2} + \frac{\nu}{2}\right) {}_1F_1\left(-\frac{\nu}{2}, \frac{1}{2}, \eta^2\right) \quad (23a)$$

$$g_\nu^{(o)}(\eta) \equiv \pi^{-1/2} \Gamma\left(1 + \frac{\nu}{2}\right) \eta {}_1F_1\left(\frac{1}{2} - \frac{\nu}{2}, \frac{3}{2}, \eta^2\right) \quad (23b)$$

<sup>†</sup> For convenience, we have multiplied the functions in Ref. 4 by  $2^\nu$  to obtain the expressions in Equation (21).

whence from Eqs. (21a) and (21b):

$$H_\nu(\eta) = \cos \frac{\pi \nu}{2} g_\nu^{(e)}(\eta) + \sin \frac{\pi \nu}{2} g_\nu^{(o)}(\eta) \quad (24a)$$

$$h_\nu(\eta) = -\sin \frac{\pi \nu}{2} g_\nu^{(e)}(\eta) + \cos \frac{\pi \nu}{2} g_\nu^{(o)}(\eta) \quad (24b)$$

Conversely,

$$g_\nu^{(e)}(\eta) = \cos \frac{\pi \nu}{2} H_\nu(\eta) - \sin \frac{\pi \nu}{2} h_\nu(\eta) \quad (25a)$$

$$g_\nu^{(o)}(\eta) = \sin \frac{\pi \nu}{2} H_\nu(\eta) + \cos \frac{\pi \nu}{2} h_\nu(\eta) \quad (25b)$$

The modal fields in the bounded core region are therefore given by:

$$f_\nu^{(\xi)}(\eta) = \exp(-\frac{\eta^2}{2}) \left[ \begin{pmatrix} \cos \frac{\pi \nu}{2} \\ \sin \frac{\pi \nu}{2} \end{pmatrix} H_\nu(\eta) + \begin{pmatrix} -\sin \frac{\pi \nu}{2} \\ \cos \frac{\pi \nu}{2} \end{pmatrix} h_\nu(\eta) \right] \quad (26)$$

where the upper and lower symbols refer to solutions with even and odd  $x$ -symmetry, respectively. Hashimoto<sup>2</sup> obtained the same relation (with slightly different normalization of  $h_\nu(\eta)$ ) using a more complicated procedure based on contour deformation in integral representations of the solutions of the differential equation (16) (with  $\ell \rightarrow \nu$ ).

Equation (26) is in the canonical form Eq. (p2) whence one can identify the decay and growth functions as

$$\phi_\nu(\eta) = \exp(-\frac{\eta^2}{2}) H_\nu(\eta) \quad (27)$$

$$\psi_\nu(\eta) = \exp(-\frac{\eta^2}{2}) h_\nu(\eta) \quad (28)$$

Moreover, from Eq. (20) (with  $\ell \rightarrow \nu$ ) and Eq. (10),

$$\alpha_\nu^2 = n_0^2 k^2 - k a_0 (2\nu + 1) - k^2 n_2^2 \quad (29)$$

or for the special case  $n_2 = n(x_c)$  (see Fig. 1),

$$\alpha_\nu = k a_0 x_c \left[ 1 - \frac{2\nu + 1}{k a_0 x_c^2} \right]^{1/2} \quad (30)$$

The modal eigenvalues are then obtained from Eq. (13), using therein the functions identified in Equations (27) to (30).

For comparison with results derived by the high-frequency asymptotic method to be treated elsewhere, we list here the asymptotic behavior of the solutions. From Eq. (18), the high frequency limit ( $ka_0$  large) corresponds to  $\eta \gg 1$  provided that  $x \neq 0$ . Alternatively,  $\eta \gg 1$  can be achieved for moderate  $ka_0$  but sufficiently large  $x$ . These restrictions are necessary to validate the asymptotic approximation. The same remarks apply to  $\eta_0$ .

The asymptotic expansions for the Hermite functions for  $\eta \gg 1$  are<sup>5</sup>

$$\phi_\nu(\eta) = \eta^\nu \exp(-\frac{\eta^2}{2}) \sum_{m=0}^{\infty} \frac{(-1)^m (-\nu)_{2m}}{m!} (2\eta)^{-2m} \quad (31a)$$

$$\psi_\nu(\eta) = \pi^{-1/2} 2^{-\nu} \Gamma(1+\nu) \eta^{-\nu-1} \exp(\frac{\eta^2}{2}) \sum_{m=0}^{\infty} \frac{(\nu+1)_{2m}}{m!} (2\eta)^{-2m} \quad (32)$$

One may expand  $\beta_\nu$  from Eq. (20), with  $\ell \rightarrow \nu$ , in inverse powers of  $k$ :

$$\beta_\nu \sim n_0 k \left[ 1 - \frac{a_0(2\nu+1)}{2n_0^2 k} - \frac{a_0^2(2\nu+1)^2}{8n_0^4 k^2} + O(\frac{1}{k^3}) \right] \quad (33a)$$

but we observe that  $\beta_\nu^2$  is given exactly by

$$\beta_\nu^2 = n_0^2 k^2 \left[ 1 - \frac{a_0(2\nu+1)}{kn_0^2} \right] \quad (33b)$$

For perturbed parabolic profiles, it is therefore preferable to obtain the asymptotic expansion for  $\beta_\nu^2$ , and to define  $\beta_\nu$  as  $(\beta_\nu^2)^{1/2}$ .

#### D. Hyperbolic Tangent and Secant Profiles

The hyperbolic tangent profile (see Fig. 2) is defined as

$$n^2(x) = \begin{cases} n_0^2 - a_0^2 \tanh^2(bx) & |x| \leq x_c \\ n_2^2 & |x| > x_c \end{cases} \quad (34)$$

where  $n_0 \equiv n(0)$  and  $a_0, b, n_2$  are the profile constants. Therefore, Eq. (5) becomes

$$\left[ \frac{d^2}{dx^2} + k^2 n_0^2 - a_0^2 \tanh^2(bx) - \beta_\nu^2 \right] f_\nu(x) = 0 \quad (35)$$

With the change of variable

$$\xi \equiv \tanh(bx), \quad \xi_c = \tanh(bx_c) \quad (36)$$

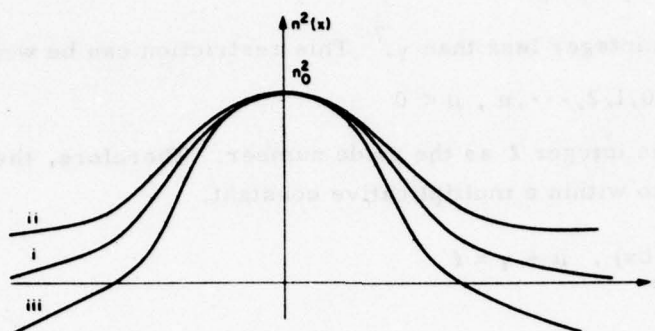


Fig. 2(a). Unbounded hyperbolic tangent profile.

- (i) hyperbolic secant profile  $a_0 = n_0$
- (ii)  $a_0 < n_0$
- (iii)  $a_0 > n_0$

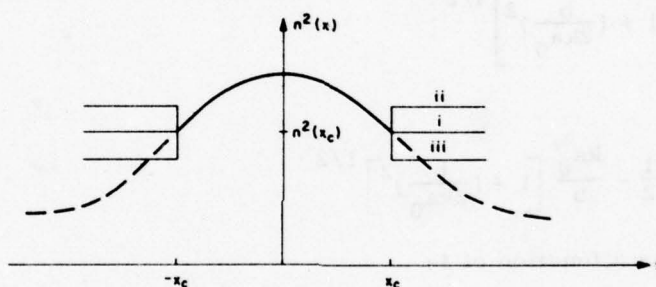


Fig. 2(b). Bounded hyperbolic tangent profile with  $a_0 < n_0$ .

- (i)  $n_2 = n(x_c)$
- (ii)  $n_2 > n(x_c)$
- (iii)  $n_2 < n(x_c)$

and the parameters

$$\gamma(\gamma + 1) = (ka_0/b)^2, \quad \mu^2 = (k^2/b^2) \left[ (\beta_v/k)^2 + a_0^2 - n_0^2 \right], \quad (37)$$

Eq. (35) takes the form of the Legendre differential equation<sup>6</sup> of degree  $\gamma$  and order  $\mu$

$$\left[ (1 - \zeta^2) \frac{d^2}{d\zeta^2} - 2\zeta \frac{d}{d\zeta} + \gamma(\gamma + 1) - \frac{\mu^2}{1 - \zeta^2} \right] f_\nu(\zeta) = 0 \quad (38)$$

Two linearly independent solutions are the Legendre functions  $P_Y^\mu(\zeta)$  and  $Q_Y^\mu(\zeta)$ .

For the uncladded profile ( $x_c \rightarrow \infty$ ), the bounded solution as  $|x| \rightarrow \infty$  ( $\zeta \rightarrow \pm 1$ ) is  $P_Y^\mu(\zeta)$  provided that  $\mu$  takes on negative values

$$\mu = -\gamma, -\gamma+1, \dots, -\gamma+n \quad (39)$$

where  $n$  is the largest integer less than  $\gamma$ .<sup>7</sup> This restriction can be written as

$$\mu + \gamma = \ell; \ell = 0, 1, 2, \dots, n, \mu < 0 \quad (40)$$

and one can identify the integer  $\ell$  as the mode number. Therefore, the solution for the uncladded profile is, to within a multiplicative constant,

$$f_{\ell}(x) \propto P_{\gamma}^{\mu}(\tanh bx), \quad \mu + \gamma = \ell \quad (41)$$

while from Eq. (37),

$$\beta_{\ell}^2(\mu) = b^2 \mu^2 \left[ 1 + (n_0^2 - a_0^2) \left( \frac{k}{b\mu} \right)^2 \right] \quad (42)$$

Since from Eq. (37),

$$\gamma = -\frac{1}{2} + \frac{ka_0}{b} \left[ 1 + \left( \frac{b}{2ka_0} \right)^2 \right]^{1/2} \quad (43)$$

and from Eq. (40)

$$\mu = \ell - \gamma = \ell + \frac{1}{2} - \frac{ka_0}{b} \left[ 1 + \left( \frac{b}{2ka_0} \right)^2 \right]^{1/2} \quad (44)$$

one may express  $\beta_{\ell}$  as a function of  $\ell$ :

$$\beta_{\ell}^2 = k^2 \left\{ n_0^2 - \frac{ba_0}{k} (2\ell + 1) \left[ 1 + \left( \frac{b}{2ka_0} \right)^2 \right]^{1/2} + \left( \frac{b}{k} \right)^2 \left( \ell^2 + \ell + \frac{1}{2} \right) \right\} \quad (45)$$

For the cladded profile,  $\ell$  is replaced by  $\nu$  in Eq. (44) and Eq. (45) and the solution  $f_{\nu}(\zeta)$  of Eq. (38) is written as a linear combination of  $P_{\nu}^{\mu}(\zeta)$  and  $Q_{\nu}^{\mu}(\zeta)$ , with

$$\mu + \gamma = \nu \quad (46)$$

The Legendre functions can be represented in terms of the hypergeometric function  ${}_2F_1$  in the form<sup>8†</sup>

---

<sup>†</sup> In order to match the solution of Eq. (47) to the solution Eq. (21) for the parabolic profile at  $x=0$ , the representations for  $P_{\nu}^{\mu}(\zeta)$  and  $Q_{\nu}^{\mu}(\zeta)$  in Ref. 8 are renormalized by multiplying by  $c_1$  and  $c_2$  respectively, where  $c_1 = 2^{-\mu} \Gamma(1 + \frac{\nu}{2} - \mu)$  and  $c_2 = \frac{2}{\pi} c_1$ . This yields Equation (47). There is a misprint in Ref. 8: the definition of  $Q_{\nu}^{\mu}(\zeta)$  there should be corrected by multiplying the first term by  $(-x)$ .

$$P_Y^\mu(\zeta) = \sqrt{\pi} (1 - \zeta^2)^{-\mu/2} \left\{ \left[ \Gamma\left(\frac{1}{2} - \frac{\nu}{2}\right) \right]^{-1} F_1 - 2\zeta \Gamma\left(1 + \frac{\nu}{2} - \mu\right) \cdot \left[ \Gamma\left(\frac{1}{2} + \frac{\nu}{2} - \mu\right) \Gamma\left(-\frac{\nu}{2}\right) \right]^{-1} F_2 \right\} \quad (47a)$$

$$Q_Y^\mu(\zeta) = \frac{\pi}{\pi} (1 - \zeta^2)^{-\mu/2} \left\{ -\tan \frac{\pi\nu}{2} \left[ \Gamma\left(\frac{1}{2} - \frac{\nu}{2}\right) \right]^{-1} F_1 - 2\zeta \cos \frac{\pi\nu}{2} \Gamma\left(1 + \frac{\nu}{2} - \mu\right) \cdot \left[ \Gamma\left(\frac{1}{2} + \frac{\nu}{2} - \mu\right) \Gamma\left(-\frac{\nu}{2}\right) \right]^{-1} F_2 \right\} \quad (47b)$$

where

$$F_1 \equiv {}_2F_1\left(-\frac{\nu}{2}, \frac{1}{2} + \frac{\nu}{2} - \mu, \frac{1}{2}, \zeta^2\right), \quad F_2 \equiv {}_2F_1\left(\frac{1}{2} - \frac{\nu}{2}, 1 + \frac{\nu}{2} - \mu, \frac{3}{2}, \zeta^2\right) \quad (47c)$$

and the hypergeometric function  ${}_2F_1$  is defined as

$${}_2F_1(a, b, c, z) = \sum_{m=0}^{\infty} \frac{(a)_m (b)_m}{m! (c)_m} z^m; \quad (48)$$

$$(a)_m = a(a+1) \cdots (a+m-1) = \frac{\Gamma(a+m)}{\Gamma(a)}$$

For notational simplicity, we retain the parameters  $\mu$ ,  $\nu$  and  $\gamma$  although  $\mu$  can be expressed in terms of  $\nu$  via Eq. (44), with  $\ell \rightarrow \nu$ , and  $\gamma = \nu - \mu$ .

Solutions  $f_\nu^{(e)}(\zeta)$  and  $f_\nu^{(o)}(\zeta)$  with even and odd  $\zeta$ -symmetry can now be defined as

$$f_\nu^{(e)}(\zeta) = \sqrt{\pi} (1 - \zeta^2)^{-\mu/2} \left[ \Gamma\left(\frac{1}{2} - \frac{\nu}{2}\right) \cos \frac{\pi\nu}{2} \right]^{-1} F_1 = \pi^{-1/2} \Gamma\left(\frac{1}{2} + \frac{\nu}{2}\right) (1 - \zeta^2)^{-\mu/2} F_1 \quad (49a)$$

$$\begin{aligned} f_\nu^{(o)}(\zeta) &= -2\sqrt{\pi} (1 - \zeta^2)^{-\mu/2} \Gamma\left(1 + \frac{\nu}{2} - \mu\right) \left[ \sin \frac{\pi\nu}{2} \Gamma\left(\frac{1}{2} + \frac{\nu}{2} - \mu\right) \Gamma\left(-\frac{\nu}{2}\right) \right]^{-1} \zeta F_2 \\ &= 2\pi^{-1/2} \Gamma\left(1 + \frac{\nu}{2}\right) \left[ \Gamma\left(1 + \frac{\nu}{2} - \mu\right) / \Gamma\left(\frac{1}{2} + \frac{\nu}{2} - \mu\right) \right] (1 - \zeta^2)^{-\mu/2} \zeta F_2 \end{aligned} \quad (49b)$$

whence from Eqs. (47a) and (47b)

$$P_Y^\mu(\zeta) = \cos \frac{\pi\nu}{2} f_\nu^{(e)}(\zeta) + \sin \frac{\pi\nu}{2} f_\nu^{(o)}(\zeta) \quad (50a)$$

$$Q_Y^\mu(\zeta) = -\sin \frac{\pi\nu}{2} f_\nu^{(e)}(\zeta) + \cos \frac{\pi\nu}{2} f_\nu^{(o)}(\zeta) \quad (50b)$$

Conversely,

$$f_\nu^{(e)}(\zeta) = \cos \frac{\pi\nu}{2} P_Y^\mu(\zeta) - \sin \frac{\pi\nu}{2} Q_Y^\mu(\zeta) \quad (51a)$$

$$f_\nu^{(o)}(\zeta) = \sin \frac{\pi\nu}{2} P_Y^\mu(\zeta) + \cos \frac{\pi\nu}{2} Q_Y^\mu(\zeta) \quad (51b)$$

which can be combined into the canonical form Eq. (12):

$$f_v^{(s)}(\zeta) = \begin{Bmatrix} \cos \frac{\pi v}{2} \\ \sin \frac{\pi v}{2} \end{Bmatrix} P_Y^\mu(\zeta) + \begin{Bmatrix} -\sin \frac{\pi v}{2} \\ \cos \frac{\pi v}{2} \end{Bmatrix} Q_Y(\zeta) \quad (52)$$

Thus, referring to Eq. (12), one identifies the decaying and growing solutions as

$$\phi_v(x) = P_Y^\mu(\tanh bx) \quad (53a)$$

$$\psi_v(x) = Q_Y^\mu(\tanh bx) \quad (53b)$$

The modal eigenvalues are inferred from Eqs. (13) and (53), where from Eq. (10) and from Eq. (45), with  $l \rightarrow v$ ,

$$\alpha_v = k \left\{ n_0^2 - n_2^2 - \frac{ba_0}{k} (2v+1) \left[ 1 + \left( \frac{b}{2ka_0} \right)^2 \right]^{1/2} + \left( \frac{b}{k} \right)^2 (v^2 + v + \frac{1}{2}) \right\}^{1/2} \quad (54a)$$

For the special case  $n_2 = n(x_c)$

$$\alpha_v = ka_0 \zeta_c \left\{ 1 - \frac{b}{ka_0 \zeta_c^2} (2v+1) \left[ 1 + \left( \frac{b}{2ka_0} \right)^2 \right]^{1/2} + \left( \frac{b}{ka_0 \zeta_c^2} \right)^2 \zeta_c^2 (v^2 + v + \frac{1}{2}) \right\}^{1/2} \quad (54b)$$

The representations Eq. (47) for the Legendre functions are not appropriate for the high frequency range because the parameter  $\mu$ , which is proportional to  $k$  (see Eq. (44)), appears in the numerator of the series expansion Eq. (48) for  $F_1$  and  $F_2$ . Instead, another representation, which has  $\mu$  in the denominator of the series expansion, has been selected from the variety listed in Ref. 9 (this representation is normalized in accordance with the footnote on page 132):

$$P_Y^\mu(\zeta) = \left[ \Gamma(1 + \frac{v}{2} - \mu) \Gamma(1 - \mu) \right] \zeta^v (1 - \zeta^2)^{-\mu/2} {}_2F_1(-\frac{v}{2}, \frac{1}{2} - \frac{v}{2}, 1 - \mu, 1 - \zeta^{-2}) \quad (55a)$$

$$Q_Y^\mu(\zeta) = 2^{-v} \pi^{-1/2} \Gamma(1 + v) \left[ \Gamma(1 + \frac{v}{2} - \mu) / \Gamma(\frac{3}{2} + v - \mu) \right] \zeta^{-v-1} (1 - \zeta^2)^{\mu/2} {}_2F_1(1 + \frac{v}{2}, \frac{1}{2} + \frac{v}{2}, \frac{3}{2} + v - \mu, \zeta^{-2}) \quad (55b)$$

The corresponding asymptotic expansion for  $\beta_v^2$  is from Eq. (45), with  $l \rightarrow v$ :

$$\beta_v^2 = k^2 \left[ n_0^2 - (2v+1) \frac{ba_0}{k} + \frac{b^2}{k^2} (v^2 + v + \frac{1}{2}) - \frac{(2v+1)b^3}{8a_0k^3} + O(k^{-5}) \right] \quad (56)$$

For special case  $a_0 = n_0$ , the hyperbolic tangent profile becomes the hyperbolic secant

$$n^2(x) = n_0^2 \operatorname{sech}^2(bx) \quad (57)$$

which has the important property that all of its guided modes travel with the same group velocity,  $v_g$ . From Eq. (42),  $\beta_v = b\mu$  or

$$\beta_v = bv + \frac{b}{2} - kn_0 \left[ 1 + \left( \frac{b}{2kn_0} \right)^2 \right]^{1/2} \quad (58)$$

whence, the group velocity, defined by  $v_g = \frac{1}{c_0} \frac{d\beta_v}{dk}$ , is independent of  $v$ .

The modal solutions for the hyperbolic secant profile can be determined from Eqs. (55), (54), (58) by substituting  $a_0 = n_0$ .

Joint Services Technical Advisory Committee  
F44620-74-C-0056

E. Navon and L. B. Felsen

National Science Foundation  
ENG 75-22625

#### REFERENCES

1. S. Choudhary and L. B. Felsen, "Asymptotic Theory of Ducted Propagation," J. Acoust. Soc. Am., Vol. 63, No. 3, pp. 661-666 (March 1978).
2. M. Hashimoto, "The Effect of an Outer Layer on Propagation in a Parabolic Index Optical Waveguide," Inst. J. of Elect., Vol. 39, No. 5, pp. 579-582 (1975).
3. N. N. Lebedev, "Special Functions and Their Applications," (Englewood Cliffs, N.J.: Prentice Hall, 1965) p. 284.
4. "Theory and Applications of Special Functions," Ed. R. A. Askey, Academic Press, p. 369 (1975).
5. N. N. Lebedev, *ibid.*, pp. 291-293.
6. W. Magnus, F. Oberhettinger and R. J. Soni, "Formula and Theorems of the Special Functions of Mathematical Physics," Springer-Verlag, N. Y., p. 151 (1966).
7. E. T. Kornhauser and A. D. Yaghjian, "Modal Solution of a Point Source on a Strongly Focusing Medium," Radio Science, Vol. 2 (New Series), No. 3, pp. 299-312 (March 1967).
8. W. Magnus, et al., *ibid.*, p. 167.
9. W. Magnus, et al., *ibid.*, p. 155 (use formula 11 and 28 with p. 166).

# GUIDED MODES IN A GRADED INDEX OPTICAL FIBER SURROUNDED BY A HOMOGENEOUS CLADDING

E. Navon and L. B. Felsen

## A. Introduction

Modal field solutions for graded index optical fibers surrounded by a homogeneous cladding must generally be obtained by approximate methods suitable for the high frequency range. To check the validity of these methods, it is important to have special exact solutions available for comparison. Here, such a solution is developed for fibers with a parabolic index core. The formulation, carried out in a manner that can be adapted to the construction of asymptotic solutions by the evanescent wave tracking method, is analogous to that described for slab waveguides elsewhere in this report.<sup>1</sup>

## B. General Formulation

We seek angularly periodic, bounded solutions  $\underline{E}(r, \theta, z)$  and  $\underline{H}(r, \theta, z)$  of the vector wave equation in cylindrical coordinates. The refractive index  $n(r)$  is assumed to be longitudinally and angularly independent. It may describe an inhomogeneous core of radius  $r_c$  surrounded by an infinite homogeneous cladding (see Figure 1)

$$n^2(r) = \begin{cases} n_0^2 - \bar{n}^2(r) & 0 \leq r \leq r_c \\ n_2^2 & r_c < r \end{cases} \quad (1)$$

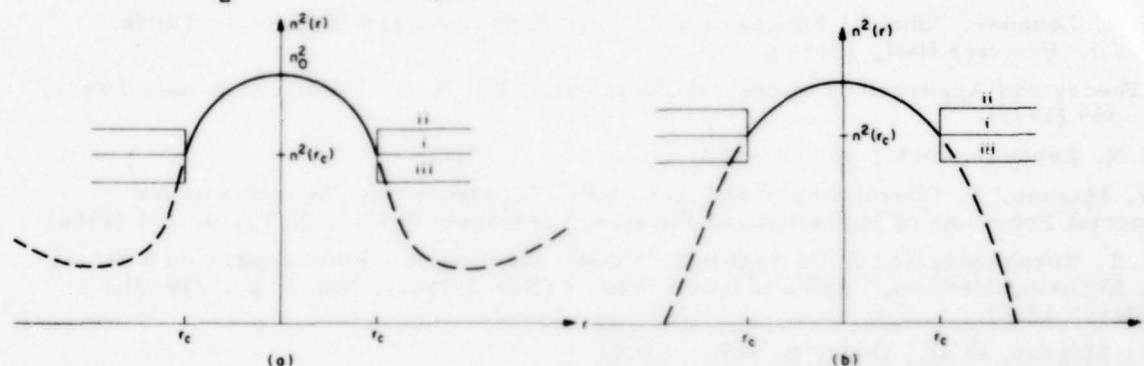


Fig. 1. Various profile shapes. (a) General symmetrical profile; (b) parabolic profile. --- unbounded profile — bounded profile with (i) continuous  $n_2 = n(r_c)$ ; (ii)  $n_2 > n(r_c)$ ; (iii)  $n_2 < n(r_c)$ .

where  $n_0 \equiv n(0)$ ,  $\bar{n}(r)$  is a given analytic function of  $r$ , and  $n_2$  is a constant.

When solving the three-dimensional vector wave equation in cylindrical coordinates by the method of separation of variables, one gets two separation parameters  $n^\dagger$  and  $\mu$  for the angular and radial domains respectively. The periodic angular dependence is of the form  $\exp(in\theta)$ , where  $n$  takes on integral values, which may be taken as positive or zero, without loss of generality:  $n = 0, 1, 2, \dots$ . The radial parameter  $\mu$  for modal fields bounded at infinity in an uncladded profile can be identified with the integers  $\mu = m = 0, 1,$

<sup>†</sup> One must distinguish between the variable refractive index  $n(r)$  and the angular parameter  $n$ .

2, ... . In the core region  $r \leq r_c$  of a cladded profile, boundedness at infinity is not required whence  $\mu$  can assume non-integer real values:

$$\mu = m + \Delta\mu_m \quad (2)$$

The cladding effect is thus expressed via  $\Delta\mu_m$  which is determined from the dispersion equation.

Kurtz and Streifer<sup>2</sup>, and later Hashimoto<sup>3</sup> and Ikuno<sup>4</sup> found that for a weakly inhomogeneous medium ( $\bar{n}(r) < n_0$  for  $r < r_c$ ) the vector wave equation can be reduced to a second order, ordinary differential equation for a scalar wave function  $R_{\mu, q}^i(r)$  of the form

$$\left[ \frac{d^2}{dr^2} + \frac{1}{r} \frac{d}{dr} + k^2 n^2(r) - \beta_{\mu, q}^2 - \frac{q^2}{r^2} \right] R_{\mu, q}^i(r) = 0 \quad (3)$$

with

$$q = |1 \mp n| \quad i = 1, 2 \quad (4)$$

where an  $\exp(in\theta + i\beta_{\mu, q} z)$  dependence is assumed and the upper (lower) sign correspond to  $i = 1$  ( $i = 2$ ). The solution of (3) can be expressed as

$$R_{\mu, q}^i(r) = A_{\mu, q}^i \phi_{\mu, q}^i(r) + D_{\mu, q}^i \psi_{\mu, q}^i(r) \quad i = 1, 2 \quad (5)$$

where, without loss of generality, the linearly independent functions  $\phi(r)$  and  $\psi(r)$  are assumed to be decaying and growing, respectively, as  $r \rightarrow \infty$ .

#### 1. Uncladded Profile

The uncladded profile is of the form

$$n^2(r) = n_0^2 - \bar{n}^2(r) \quad 0 \leq r < \infty \quad (6)$$

for which the inhomogeneous core extends to infinity and the solution (5) should hold for  $0 \leq r < \infty$ . Boundedness at  $r \rightarrow \infty$  excludes  $\psi_{\mu, q}(r)$  from the solution. Moreover, it is assumed that the function  $\phi_{\mu, q}(r)$  is now described by integral values of  $\mu = m = 0, 1, 2, \dots$  so that the  $r$  dependent part of the solution for the uncladded profile is

$$R_{\mu, q}^i(r) = R_{m, q}^i(r) = A_{m, q} \phi_{m, q}^i(r) \quad i = 1, 2 \quad (7)$$

#### 2. Cladded Profile

For the cladded profile Eq. (1) the solution in the core region is  $R_{\mu, q}^i(r)$  (see Eq. (5)), and in the outer region the solution is assumed in the form

$$R_{\mu, q}^i(r_c) \left[ K_n(\alpha_{\mu, q} r) / K_n(\alpha_{\mu, q} r_c) \right], \quad r > r_c \quad i = 1, 2 \quad (8)$$

where  $K_n(w)$  is the modified Bessel function of  $n$ -th order and  $\alpha_{\mu, q}$  is the transverse propagation coefficient defined as

$$\alpha_{\mu, q} = \left[ \beta_{\mu, q}^2 - k_{n_2}^2 \right]^{1/2} \quad (9)$$

At the boundary  $r = r_c$  the required continuity of the tangential electric field is already satisfied by the assumed form of Equation (8). Continuity of the tangential magnetic field requires

$$\frac{d}{dr} R_{\mu, q}^i(r_c) + \bar{\alpha}_{\mu, q} R_{\mu, q}^i(r_c) = 0 \quad i = 1, 2 \quad (10)$$

where for the special case  $n_2 = n(r_c)$  it can be shown that

$$\bar{\alpha}_{\mu, q} = \frac{1}{r_c} \bar{\tau}_n + \frac{\alpha_{\mu, q}^2 r_c}{K_n'(w) \Big|_{w=r_c} - \bar{\tau}_n K_n(w) \Big|_{w=r_c}} \quad i = 1, 2 \quad (11)$$

and the prime denotes differentiation with respect to the argument. Normalizations will be sought such that the  $r$ -dependent part of the solution  $R_{\mu, q}^i(r)$  can be cast into the canonical form

$$R_{\mu, q}^i(r) = \cos \pi \mu \phi_{\mu, q}^i(r) - \sin \pi \mu \psi_{\mu, q}^i(r) \quad i = 1, 2 \quad (12)$$

which,  $\mu \rightarrow m$ , reduces to the uncladded profile solution Equation (7). For the parabolic refractive index, the form in Eq. (12) was also obtained by Hashimoto<sup>3</sup> via manipulation of contour integral representations of the solution for the unbounded profile. We shall use an alternative procedure based directly on known properties of the confluent hypergeometric functions.

From Eq. (12) and the dispersion equation (10), one obtains the transcendental equation for  $\Delta \mu_m$  in Eq. (2),

$$\tan \pi \mu = \tan(\pi \Delta \mu_m) = \frac{\phi'(r_c) + \bar{\alpha} \phi(r_c)}{\psi'(r_c) + \bar{\alpha} \psi(r_c)} \quad i = 1, 2 \quad (13)$$

where, as a matter of convenience, we omitted the subscripts  $\mu$  and  $q$ ; the prime denotes differentiation with respect to  $r$  and  $\bar{\alpha}$  and is defined in Equation (11).

### C. Parabolic Profile

The parabolic profile is defined as

$$n^2(r) = \begin{cases} n_0^2 - a_0^2 r^2 & r \leq r_c \\ n_2^2 & r > r_c \end{cases} \quad (14)$$

The solution for the uncladded case  $r_c \rightarrow \infty$  is known to be<sup>2</sup>

$$R_{m,q}^i = r^q \exp(-ka_0 r^2/2) L_m^{(q)}(ka_0 r^2) \quad i = 1, 2 \quad (15)$$

also

$$\beta_{m,q}^2 = k^2 n_0^2 \left[ 1 - \frac{2a_0(2m+q+1)}{kn_0^2} \right] \quad i = 1, 2 \quad (16)$$

where  $q$  and  $m$  are the modified angular (see Eq. (4)) and radial mode indices, respectively, and  $L_m^{(q)}$  is the generalized Laguerre polynomial.

For the core region of the cladded profile,  $m$  is replaced by  $\mu$  as in Eq. (2), and the Laguerre function  $L_\mu^{(q)}$  is expressed via the confluent hypergeometric function  ${}_1F_1$ :<sup>5</sup>

$$L_\mu^{(q)}(\xi^2) = \frac{\Gamma(1+\mu+q)}{\Gamma(1+\mu)\Gamma(1+q)} {}_1F_1(-\mu, 1+q, \xi^2) \quad (17)$$

where

$$\xi^2 \equiv ka_0 r^2, \quad \xi_c^2 = ka_0 r_c^2 \quad (17a)$$

Decomposition of the solution into the decaying and growing functions of the canonical form Eq. (12) can be carried out by use of the connection formula between the first ( ${}_1F_1$ ) and second ( $U$ ) confluent hypergeometric functions<sup>6</sup>

$$\begin{aligned} {}_1F_1(-\mu, 1+q, \xi^2) &= \frac{\Gamma(1+q)}{\Gamma(1+q+\mu)} \cos \pi \mu U(-\mu, 1+q, \xi^2) \\ &+ \frac{\Gamma(1+q)}{\Gamma(-\mu)} \exp(+\xi^2) U(1+\mu+q, 1+q, -\xi^2) \end{aligned} \quad (18a)$$

where

$$U(a, c, w) = \frac{\pi}{\sin(\pi c)} \left[ \Gamma(c) \Gamma(1+a-c) \right]^{-1} {}_1F_1(a, c, w) \\ - w^{1-c} \left[ \Gamma(a) \Gamma(2-c) \right]^{-1} {}_1F_1(a+1-c, 2-c, w) \quad (18b)$$

which with Eq. (17) yields

$$L_{\mu}^{(q)}(\xi^2) = \cos \pi \mu \left[ \Gamma(1+\mu) \right]^{-1} U(-\mu, 1+q, \xi^2) \\ - \sin(\pi \mu) \pi^{-1} \Gamma(1+\mu+q) \exp(\xi^2) U(1+\mu+q, 1+q, -\xi^2) \quad (19)$$

Comparison with Eq. (12) gives

$$\phi_{\mu, q}^i(r) = r^q \exp(-ka_0 r^2/2) \left[ \Gamma(1+\mu) \right]^{-1} U(-\mu, 1+q, ka_0 r^2) \quad (20a)$$

$$\psi_{\mu, q}^i = r^q \exp(ka_0 r^2/2) \Gamma(1+\mu+q) \pi^{-1} U(1+\mu+q, 1+q, -ka_0 r^2) \quad (20b)$$

The propagation coefficients are:

$$\beta_{\mu, q}^2 = k^2 n_0^2 \left[ 1 - \frac{2a_0}{kn_0^2} (2m+q+1+2\Delta\mu_m) \right] \quad (21)$$

where  $\Delta\mu_m$  is to be determined from Eq. (13), using the functions in Equation (20).

The asymptotic expansion of the confluent hypergeometric function<sup>7</sup> for  $w \rightarrow \infty$ ,  $-\frac{3\pi}{2} < \arg w < \frac{3\pi}{2}$  is

$$U(a, c, w) = w^{-a} \sum_{j=0}^J (-1)^j \frac{(a)_j (a-c+1)_j}{j! w^j} + O(|w|^{-a-1-J}), \quad j = 0, 1, 2, \dots \quad (22)$$

Thus, for  $\xi^2 = ka_0 r^2 \rightarrow \infty$

$$\phi_{\mu, q}^i(\xi) = \exp(-\xi^2/2) \xi^{2\mu+q} \left[ \Gamma(1+\mu) \right]^{-1} \sum_{j=0}^J (-1)^j \frac{(-\mu)_j (-\mu-q)_j}{j!} \xi^{-2j} \quad (23a)$$

$$\psi_{\mu, q}^i(\xi) = \exp(\xi^2/2) \xi^{-2\mu-q-2} \pi^{-1} \Gamma(1+\mu+q) \sum_{j=0}^J \frac{(1+\mu+q)_j (1+\mu)_j}{j!} \xi^{-2j} \quad (23b)$$

For the Bessel functions, one has<sup>8</sup>

$$-wK'_n(w)/K_n(w) = w \left[ 1 + \frac{1}{2w} + \frac{4n^2-1}{8w^2} + O(w^{-3}) \right] \quad (24)$$

It is convenient to introduce a "combined mode parameter"  $\nu$  instead of  $\mu$

$$\nu \equiv 2\mu + q \quad (25)$$

With this change the asymptotic expressions (23) become

$$\begin{aligned} \phi_{\nu, q}^i(\xi) = \exp(-\xi^2/2) \xi^\nu \left[ \Gamma\left(\frac{\nu-q+2}{2}\right) \right]^{-1} \left\{ 1 - \frac{\nu^2-q^2}{4\xi^2} \right. \\ \left. + \frac{(\nu^2-q^2)[(\nu-2)^2-q^2]}{32\xi^4} + O(\xi^{-6}) \right\} \end{aligned} \quad (26a)$$

$$\begin{aligned} \psi_{\nu, q}^i(\xi) = \exp(\xi^2/2) \xi^{-\nu-2} \pi^{-1} \Gamma\left(\frac{\nu+q+2}{2}\right) \left\{ 1 + \frac{(\nu+2)^2-q^2}{4\xi^2} \right. \\ \left. + \frac{[(\nu+2)^2-q^2][(\nu+4)^2-q^2]}{32\xi^4} + O(\xi^{-6}) \right\} \end{aligned} \quad (26b)$$

Also

$$\beta_{\nu, q}^2 = k^2 n_0^2 \left[ 1 - \frac{2a_0}{kn_0} (\nu+1+\Delta\nu_m) \right] \quad (26c)$$

Moreover,  $\Delta\mu_m \rightarrow \Delta\nu_m/2$  in Equation (13).

Joint Services Technical Advisory Committee  
F44620-74-C-0056

E. Navon and L. B. Felsen

National Science Foundation  
ENG75-22625

#### REFERENCES

1. E. Navon and L.B. Felsen, "Propagation in Inhomogeneous Slot Waveguides -- Exact Solution," Progress Report No. 43 to 55 FAC, Polytechnic Institute of New York, Report No. R-452. 43-78 (1978).
2. C.N. Kurtz and W. Streifer, "Guided Waves in Inhomogeneous Focusing Media. Part I: Formulation, Solution for Quadratic Inhomogeneity," IEEE Trans. on Microwave Th. and Tech., Vol. MTT-17, No. 1, pp. 11-15, January 1969. Part II: "Asymptotic Solution for General Weak Inhomogeneity," ibid. No. 5, pp. 250-253, May 1969.
3. M. Hashimoto, "Analysis of Guided Waves Along the Cladded Optical Fibers: Parabolic Index Core and Homogeneous Cladding," IEEE Trans. Microwave Th. and Tech., Vol. MTT-25, No. 1, pp. 11-17, January 1977.

4. H. Ikuno, "Analysis of Guided Modes in a Graded-Index Optical Fiber," submitted for publication, May 1977.
5. Magnus, et al., *ibid*, p. 336.
6. Magnus, et al., *ibid*, p. 263.
7. Magnus, et al., *ibid*, p. 289.
8. Magnus, et al., *ibid*, p. 139.

## THE DIFFRACTION OF GAUSSIAN BEAMS BY PERIODIC LAYERS

R.S. Chu, J.A. Kong and T. Tamir

A. Background

Because of its relevance to holography, to the diffraction of light by sound, as well as to other applications, the scattering of a plane wave by a periodically modulated layer has been treated extensively in the past by using a variety of methods.<sup>1</sup> The simplest approach<sup>2</sup> is to replace the layer by a half space wherein the field is expressed in terms of two first-order coupled waves; however, this procedure neglects the effects of the exit boundary. For a layer bounded by identical media on both sides, the configuration is symmetric and its relevant electromagnetic problem has been solved rigorously by Chu and Tamir.<sup>3</sup> The asymmetric configuration of a layer bounded by different media on its two sides has been studied with a rigorous modal approach by Chu and Kong.<sup>4,5</sup>

If the incident field is a realistically bounded beam, the diffracted field may be quite different from that of an incident plane wave of infinite extent. For a Gaussian beam incident at or close to the Bragg angle on a modulated half space, Chu<sup>6</sup> and Chu and Tamir<sup>7,8</sup> have shown that (a) the diffracted beams are formed by depletion of energy away from the center of the original beam, (b) the transmitted beams split at the exit plane of the film, and (c) complete conversion of energy into the Bragg field cannot generally occur. Upon examining the patterns of Gaussian beams diffracted by a thick holographic film, Forshaw<sup>9</sup> verified items (a) and (b) and he observed ripples in the far-field intensity of the emerging beams. Forshaw also derived<sup>9</sup> the far-field profiles of these beams by using a Fourier transform of Chu's field solution<sup>6</sup> at the exit plane of the modulated layer. While the patterns thus obtained by Forshaw provide an extension of the half-space results, they do not predict the observed ripples in the far field pattern.

In the present work, we treat the diffraction of a Gaussian beam by modulated layers and derive solutions for both the near and the far fields by integrating over an appropriate plane-wave spectrum. The computed results agree with Forshaw's experimental observations and they account for the ripples in the diffracted beam profiles. We also confirm Chu and Tamir's conclusions<sup>7</sup> that, in contrast to previous theoretical studies which assumed an incident plane wave, complete conversion of energy into the Bragg field cannot generally occur in the case of bounded beams. The reason for this incomplete conversion is clarified by examining the spectral contents of the beam energy.

B. Solution for the Far Field

To consider the diffraction of bounded beams, we first list succinctly the pertinent

expressions for a plane wave incident on a periodic slab, as shown in Figure 1. The

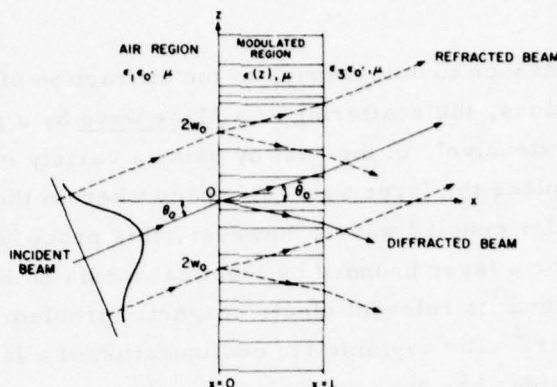


Fig. 1. Geometrical configuration of the problem.

permittivity of the slab is assumed to be

$$\epsilon(z) = \epsilon_0 \epsilon_2 [1 + M \cos Kz] \quad (1)$$

where  $K = 2\pi/\Lambda$ ,  $\epsilon_0$  is the permittivity of vacuum,  $\epsilon_2$  is the relative permittivity of the slab if  $M=0$ ,  $\Lambda$  is the periodicity, and  $M$  is the modulation index of the periodic medium. A plane wave with perpendicular polarization and characterized by an electric field vector

$$E_y = E_0 e^{ik_{1x}x + ik_{0z}z} \quad (2)$$

is assumed incident on the slab from  $z < 0$ , at an angle  $\theta$  with respect to the normal to the slab. The time dependent factor  $e^{-i\omega t}$  is assumed and suppressed in Equation (2). As given, the incident propagation wave vector is  $\vec{k} = \hat{x}k_{1x} + \hat{z}k_{0z}$ , where  $k_{0z} = k_1 \sin \theta$ ,  $k_{1x} = k_1 \cos \theta$ , and

$$k_1 = \frac{2\pi}{\lambda} \sqrt{\epsilon_1} = \omega (\mu \epsilon_1 \epsilon_0)^{1/2},$$

where  $\lambda$  is the wavelength in vacuum. We assume that the incident angle  $\theta$  is at, or near, the first Bragg angle  $\theta_B$  so that the transmitted wave consists mostly of the zeroth and the first order wave components, namely,

$$E_y = T_0 e^{ik_{3x}^a x + ik_{0z} z} + T_{-1} e^{ik_{3x}^b x + ik_{-1z} z} \quad (3)$$

where

$$k_{3x}^a = (k_3^2 - k_{0z}^2)^{1/2} \quad (4)$$

$$k_{3x}^b = (k_3^2 - k_{-1z}^2)^{1/2} \quad (5)$$

$$k_{-1z} = k_{0z} - K \quad (6)$$

and  $k_3 = \omega(\epsilon_3 \epsilon_0)^{1/2}$ . The transmission coefficients  $T_0$  and  $T_{-1}$  can be found by using the second-order coupled mode equations and by properly matching the boundary conditions at  $x=0$  and  $L$ , as given in Reference 5. Hence  $T_0$  and  $T_{-1}$  are well-known for plane waves.

While this work has examined the transmitted field of an incident Gaussian beam in both the near and the far field, we present here only the latter field in order to compare it with Forshaw's observation. For this purpose, we use the Fraunhofer approximation

$$\bar{E}_y = \sqrt{\frac{k_3}{i2\pi(x-L)}} e^{ik_3x + ik_3z^2/2x} \int_{-\infty}^{\infty} dz' E_{ap}(z') e^{-ik_3z' \sin \theta} \quad (7)$$

where now  $\bar{E}_y$  denotes a field produced by a bounded beam rather than that due to an incident wave, while  $E_{ap}(z')$  is the beam field at  $x=L$ . Thus,  $\bar{E}_y$  is the Fourier transform of that aperture field. For incidence near the Bragg angle,  $E_{ap}(z')$  consists of the (zero-order) refracted field and the (-1 order) Bragg field, which are given respectively by

$$E_0(z) = \int_{-\infty}^{\infty} G(k_{0z}) T_0(k_{0z}) \exp(ik_{0z}z) dk_{0z} \quad (8)$$

$$E_{-1}(z) = \int_{-\infty}^{\infty} G(k_{0z}) T_{-1}(k_{0z}) \exp(ik_{0z}z) dk_{0z} \quad (9)$$

where  $T_0(k_{0z})$  and  $T_{-1}(k_{0z})$  are known,<sup>5</sup> as noted before, and

$$G(k_{0z}) = \frac{w_0}{\sqrt{\pi}} \exp[-(k_{0z} - k_1 \sin \theta_0)^2 w_0^2] \quad (10)$$

is the transform of the incident Gaussian field, where  $4w_0$  is the beam width projected along the  $z$  axis and  $\theta_0$  is the angle of incidence, as shown in Figure 1. Thus, for incidence at exactly the first Bragg angle,  $\theta_0 = \theta_B$ , we have

$$k_1 \sin \theta_0 = \pi/\Lambda = \frac{1}{2} K \quad (11)$$

By introducing Eqs. (8) through (10) into Eq. (7), we derived the field patterns for a large number of typical situations. One such case is illustrated in Figs. 2 and 3,

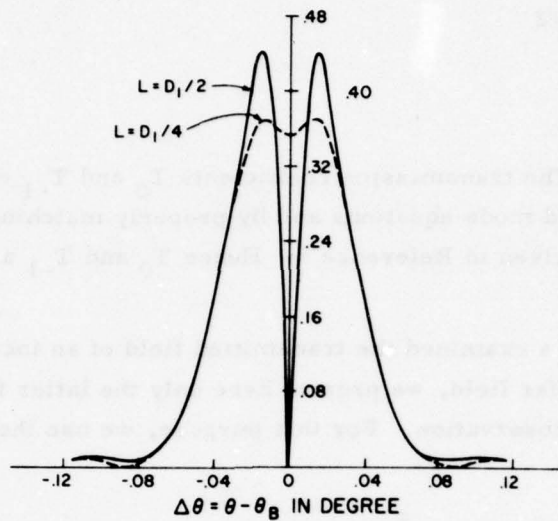


Fig. 2. Far field pattern of the zeroth order beam for  $D_1 = 8.2475 \text{ mm}$ ,  $\lambda = 0.6328 \mu\text{m}$ ,  $\Lambda = 1 \mu\text{m}$ ,  $\epsilon_1 = \epsilon_3 = 1$ ,  $\epsilon_2 = 2.25$ ,  $\overline{\omega}_0 = 500$  and  $M = 10^{-4}$ .

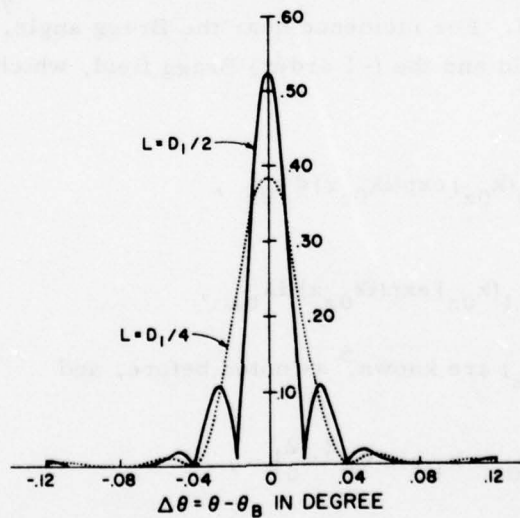


Fig. 3. Far field pattern of the Bragg-scattered beam for the same parameters as those in Figure 2.

which show the zeroth order beam and the Bragg-scattered beam, respectively. We compare the zeroth order beams for the two slab thicknesses  $L = D_1/2$  and  $L = D_1/4$ ,

as functions of  $\Delta\theta = \theta - \theta_B$  in Figure 2. Here  $D_1/2$  is the thickness for which total conversion of energy into the Bragg beam would be expected from the simple plane-wave analysis. In contrast, we see in Fig. 2 that only a deep null is produced at the center position  $\theta = \theta_B = 18.445^\circ$ , but otherwise a two-lobe zero-order beam of substantial magnitude does appear in the far field. This agrees with the analytic and experimental results obtained by Forshaw. The physical reason for this deep null is that, at  $L = D_1/2$ , the central portion of the Gaussian spectrum of the zero-order wave has completely converted its energy into the Bragg-scattered wave, which results in a depletion of energy from its beam-center position, as shown. For the case  $L = D_1/4$ , the zeroth order beam has only a small dip at  $\theta = \theta_B$  because only partial conversion of energy is expected for this slab thickness.

A comparison of the Bragg-scattered beams for the two slab thicknesses  $L = D_1/2$  and  $L = D_1/4$  is shown in Figure 3. For the case  $L = D_1/2$ , the beam width is quite narrow and side-lobe ripples are obtained. This agrees with the experimental results performed by Forshaw,<sup>9</sup> even though his analytical results have revealed no ripples for the Bragg-scattered beam. For the case  $L = D_1/4$ , the beam width becomes broadened and the side-lobe ripples are much smaller.

The foregoing considerations and results can be shown<sup>10</sup> to be consistent with the fact that, if

$$w_0 > W = \frac{\lambda^2}{2M\Lambda\epsilon_2}, \quad (12)$$

then most of the incident-beam energy is concentrated around the Bragg angle. In that case, the transmitted field behaves very much like that of a plane wave. For the case shown in Figs. 2 and 3, condition (12) is violated so that total conversion of energy into the Bragg beam does not occur.

For comparison, we show in Figs. 4 and 5 a situation that does satisfy  $w_0 > W$ , in which case nearly total conversion is obtained if  $L = D_1/2$ , as expected. Thus, the case of Figs. 4 and 5 corresponds well with the plane-wave analysis, which predicts no transmitted zero-order beam if  $L = D_1/2$  and transmitted zero-order and Bragg beams of equal magnitudes if  $L = D_1/4$ . Furthermore, ripples occur in all of these far field patterns, in agreement with Forshaw's observations.

As illustrated by the above examples, we find that our results provide an extension of previous theoretical work and yield analytical expressions consistent with the most recent experimental evidence. Further details on these aspects have been presented in a recent article,<sup>10</sup> which examines the near field as well as the far field discussed

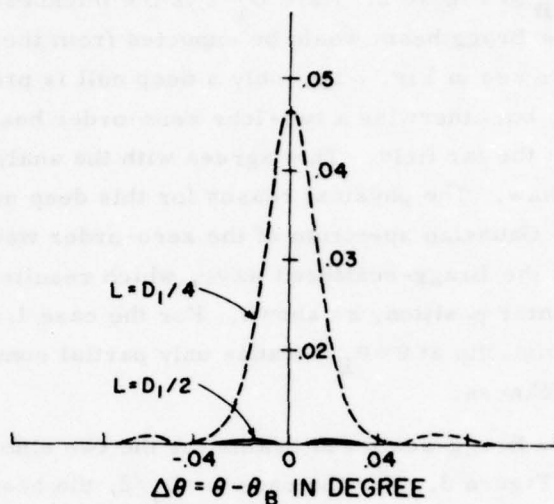


Fig. 4. Far field pattern of the zeroth order beam for  $D_1 = 0.832$  cm,  $\Lambda = 2\lambda = 1.2656 \times 10^{-4}$  cm,  $\epsilon_1 = 1$ ,  $\epsilon_2 = 2.25$ ,  $\epsilon_3 = 3$ ,  $\bar{W}_0 = 500$  and  $M = 10^{-2}$ .

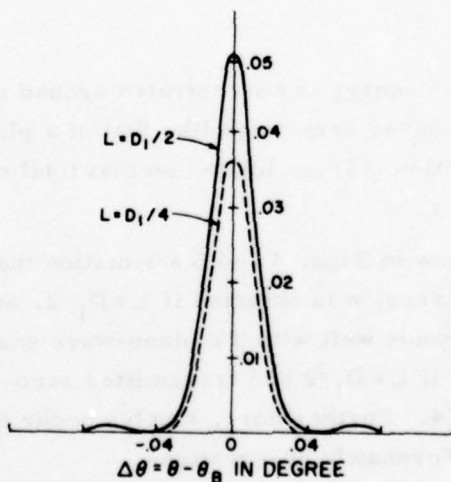


Fig. 5. Far field pattern of the Bragg-scattered beam for the same parameters as those in Figure 4.

here and also gives results for asymmetrical cases involving different exterior media, i.e.,  $\epsilon_1 \neq \epsilon_3$ .

Joint Services Technical Advisory Committee  
F44620-74-C-0056 (PINY)  
DAAB07-75-C-1346 (MIT)

T. Tamir

#### REFERENCES

1. M. Born and E. Wolf, "Principles of Optics," (New York: Pergamon Press, 1965) Ch. 12, p. 579.
2. P. Phariseau, "On the Diffraction of Light by Progressive Supersonic Waves," Proc. Ind. Acad. Sci. A, Vol. 44, pp. 165-169 (1956).
3. R.S. Chu and T. Tamir, "Guided-Wave Theory of Light Diffraction by Acoustic Microwaves," IEEE Trans, Vol. MTT-18, pp. 486-504 (1970).
4. R.S. Chu and J.A. Kong, "Modal Theory of Spatially Periodic Media," IEEE Trans. Vol. MTT-25, pp. 18-24 (1977).
5. J.A. Kong, "Second-Order Coupled Mode Equations for Spatially Periodic Media," J. Opt. Soc. Am., Vol. 67, pp. 825-829 (1977).
6. R.S. Chu, "The Diffraction of Bounded Electromagnetic Beams by Periodically Modulated Media," Polytech. Inst. of Brooklyn, Ph.D. Dissertation (1971) (University Microfilms No. 71-29042, Ann Arbor, Michigan).
7. R.S. Chu and T. Tamir, "Bragg Diffraction of Gaussian Beams by Periodically Modulated Media," J. Opt. Soc. Am., Vol. 66, pp. 220-226 (1976).
8. R.S. Chu and T. Tamir, "Diffraction of Gaussian Beams by Periodically Modulated Media for Incidence Close to a Bragg Angle," J. Opt. Soc. Am., Vol. 66, pp. 1438-1440 (1976).
9. M.R.B. Forshaw, "Diffraction of a Narrow Laser Beam by a Thick Hologram: Experimental Results," Opt. Commun., Vol. 12, pp. 279-281 (1974).
10. R.S. Chu, J.A. Kong and T. Tamir, "The Diffraction of Gaussian Beams by a Periodically Modulated Layer," J. Opt. Soc. Am., Vol. 67, pp. 1555-1561 (1977).

## BRAGG-REFLECTION APPROACH FOR BLAZED DIELECTRIC GRATINGS

K. C. Chang and T. Tamir

Optical waves supported by a dielectric grating in integrated-optics devices are often subject to considerable losses because the grating scatters energy into undesirable directions. Thus, in a beam coupler as shown in Fig. 1, an incident surface wave

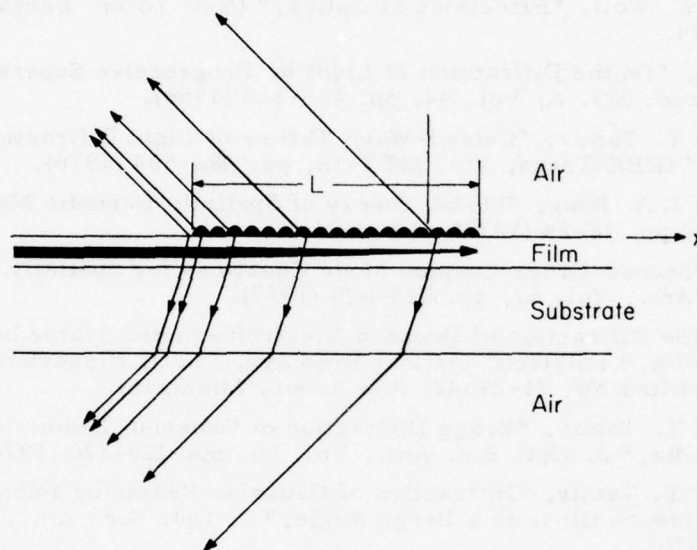


Fig. 1. Conversion of a surface wave into outgoing beams in a thin-film grating coupler.

is converted by the grating into two outgoing beams, one of which radiates into the upper air region while the second one radiates into the substrate and subsequently into the lower air region. However, usually only the former beam is utilized, so that the energy scattered into the substrate should be minimized. This can be achieved by "blazing" the grating, i.e., choosing a suitable asymmetric profile for the periodic layer, as already verified both theoretically<sup>1-3</sup> and experimentally.<sup>4</sup>

The present work addresses itself to the choice of a suitable design for achieving this blazing effect. In a previous study, Marcuse<sup>3</sup> used a geometric-optical argument for this purpose, but he pointed out that his approach becomes unwieldy and inappropriate for deep grating grooves. In the present work, we use a Bragg-reflection analysis that applies to arbitrary grating heights; preliminary results indicate that this analysis is both accurate and simple to use.

The principle of this Bragg approach is illustrated in Fig. 2, which describes a

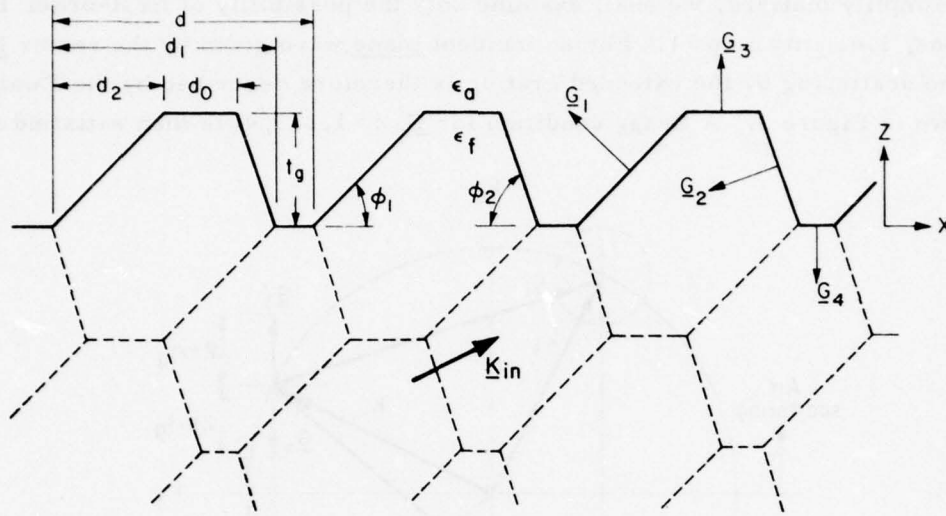


Fig. 2. Extended lattice description of a dielectric grating having a general trapezoidal profile.

rather general trapezoidal grating profile shown by the thick solid lines. Consistent with analytical results obtained by Tamir and Peng,<sup>5,6</sup> the dielectric grating exhibits the following properties:

- (1) The layer of height  $t_g$  acts as a periodic perturbation of the (volume) average dielectric constant  $\epsilon_g$  inside the layer, i.e., this constant is given here by

$$\epsilon_g = \epsilon_a + (d_0 + d_1)(\epsilon_f - \epsilon_a)/2d \quad (1)$$

- (2) The above periodic perturbation scatters an incident wave into directions determined by the grating facets, which are characterized by normal vectors  $\underline{G}_1$ ,  $\underline{G}_2$ ,  $\underline{G}_3$  and  $\underline{G}_4$  in Figure 2.

To find the directions of constructive scattering, it is then useful to extend the periodic grating configuration into the lattice structure indicated by the dashed lines. In such a lattice, Bragg-scattering theory implies that strong constructive scattering will occur along those directions that satisfy a Bragg condition with respect to any one of the following vectors:

$$\underline{G}_1 = -(2m\pi/d)(\underline{x}_0 - \underline{z}_0 \cot \phi_1)$$

$$\underline{G}_2 = -(2n\pi/d)(\underline{x}_0 + \underline{z}_0 \cot \phi_2)$$

$$\underline{G}_3 = -\underline{G}_4 = (2p\pi/t_g)\underline{z}_0$$

$$(m, n, p = \pm 1, \pm 2, \pm 3, \dots) \quad (2)$$

where  $\underline{x}_0$  and  $\underline{z}_0$  are unit vectors along the respective directions.

To simplify matters, we shall examine only the possibility of first-order Bragg interactions, i.e.,  $m=n=p=1$ . For an incident plane wave given by the vector  $\underline{K}_{in}$  in Fig. 2, the scattering by the extended grating is therefore described by the Ewald diagram shown in Figure 3. A Bragg condition for  $\underline{G}_i$  ( $i=1, 2, 3, 4$ ) is then satisfied if the

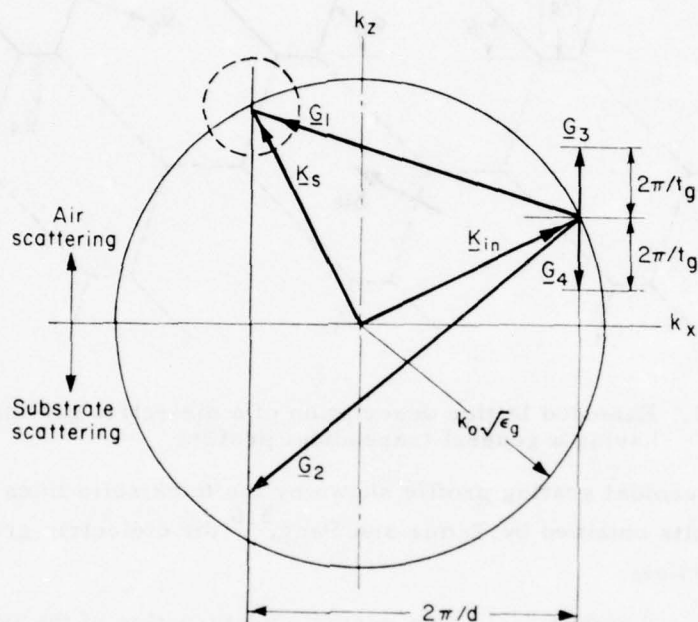


Fig. 3. Ewald diagram for scattering of a plane wave incident within the grating region shown in Figure 2.

vectorial summation of  $\underline{K}_{in}$  and  $\underline{G}_i$  yields a vector  $\underline{K}_s$  whose vertex lies on the circle of radius  $k_0\sqrt{\epsilon_g} = 2\pi\sqrt{\epsilon_g}/\lambda$ . This is the case for  $\underline{G}_1$  in Fig. 3, where strong constructive interference via Bragg reflection occurs into the upper region along the direction given by the vector  $\underline{K}_s$ . While this arrangement assures Bragg scattering with respect to  $\underline{G}_1$ , leakage along  $\underline{K}_s$  will be further enhanced if the scattering due to the other vectors  $\underline{G}_2$ ,  $\underline{G}_3$  and  $\underline{G}_4$  is appropriately reduced. The effect of  $\underline{G}_3$  and  $\underline{G}_4$  can be easily eliminated by choosing triangular (rather than trapezoidal) grating profiles. However, the vector  $\underline{G}_2$  may then still account for appreciable scattering into the substrate unless it is chosen so as to be far from satisfying a Bragg condition.

To apply the above principles to the case of a beam coupler, consider Fig. 4 in which an incident surface wave is desired to leak most of its energy into the upper (air) region. The appropriate incident wave is now shown by the arrow denoted by  $\beta$  whose

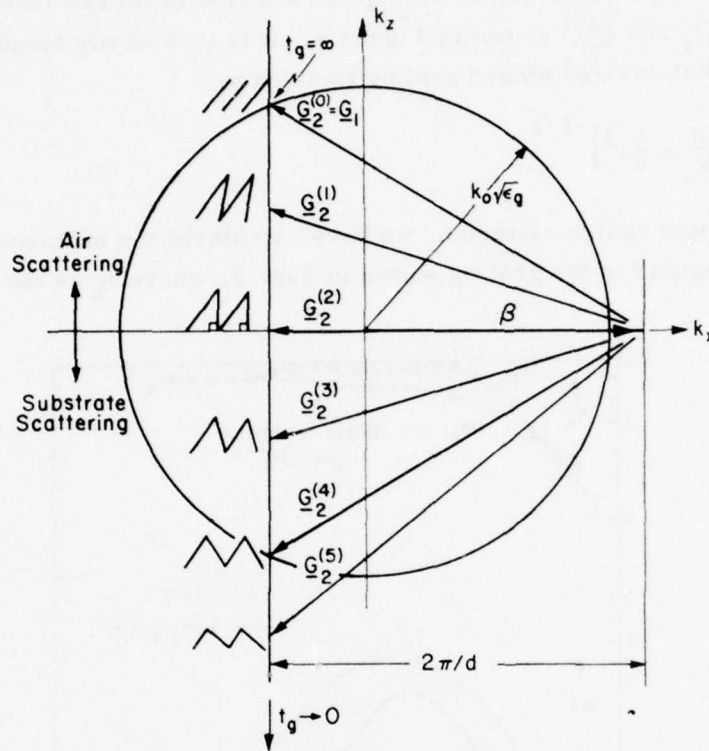


Fig. 4. Effect of varying  $\underline{G}_2$  in a triangular grating. In this case,  $\underline{G}_1$  has already been adjusted to satisfy a Bragg condition for scattering an incident surface wave into the upper (air) region.

length is larger than  $k_0\sqrt{\epsilon_g}$ . Again, we choose  $\underline{G}_1$  to produce Bragg scattering into the upper (air) region and examine possible choices for  $\underline{G}_2$  assuming now a triangular grating profile. It is then clear that, for a symmetric grating indicated by  $\underline{G}_2^{(4)}$ , a nearly equal amount of scattering should occur into the air and substrate regions, because both  $\underline{G}_1$  and  $\underline{G}_2^{(4)}$  satisfy a Bragg condition. Other typical situations for  $\underline{G}_2$  are shown in Fig. 4 by  $\underline{G}_2^{(0)}$  through  $\underline{G}_2^{(3)}$ . Of all these forms,  $\underline{G}_2^{(0)} = \underline{G}_1$  would be optimal because both  $\underline{G}_1$  and  $\underline{G}_2$  would then satisfy the same Bragg condition for scattering into the air region only. However,  $\underline{G}_2 \rightarrow \underline{G}_1$  requires very thick gratings, such as the holographic gratings used by Kogelnik and Sosnowsky.<sup>7</sup> For milled or etched gratings, on the other hand, the situation shown by  $\underline{G}_2^{(2)}$  appears to be most realistic and this corresponds to a right-angled triangle. In such a case,  $\underline{G}_2 = \underline{G}_2^{(2)}$  ends sufficiently far from the Bragg circle so that energy should be scattered primarily by  $\underline{G}_1$  only.

The preceding arguments indicate that blazing should be effectively obtained by a grating profile having a right-angled triangular shape with its two facets satisfying the construction for  $\underline{C}_1$  and  $\underline{C}_2^{(2)}$  shown in Figure 4. It is then easily found that the required relationship for that desired blazed grating is given by

$$\frac{t_g}{\lambda} = \left[ \epsilon_g - \left( \frac{\beta}{k_0} - \frac{\lambda}{d} \right)^2 \right]^{-1/2} \quad (3)$$

To verify this simple design criterion, we have calculated the efficiency  $\eta_a$  of scattering into the air region for the grating shown in Fig. 5, where  $\eta_a$  is the ratio of the

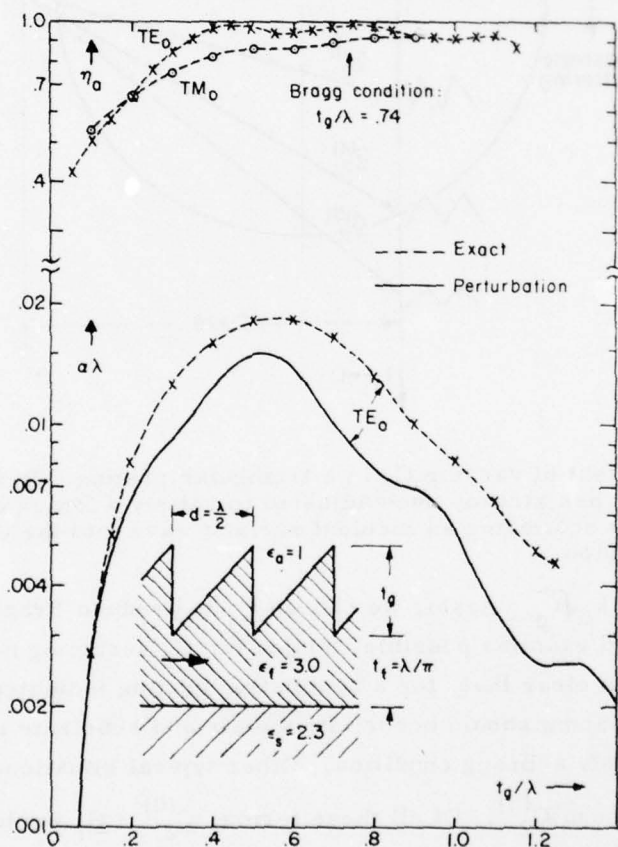


Fig. 5. Variation of efficiency  $\eta_a$  and leakage coupling  $\alpha\lambda$  in a right-angle triangular dielectric grating on which a surface wave is incident.

energy of the upper air beam and the total scattered energy. The construction prescribed by Eq. (3) predicts an optimum condition at  $t_g/\lambda = 0.74$ , as indicated by the

arrow in Figure 5. Exact calculations show that  $\eta_a$  then equals 0.95 and 0.90 for  $TE_0$  and  $TM_0$  modes, respectively. There appears, therefore, to be an excellent agreement between the simple Bragg-scattering argument for determining the grating profile and the actual result. In addition, we also note that:

- (1) Good blazing is achieved for a wide range of values, roughly  $0.5 < t_g/\lambda < 1.1$ . This is explained by the fact that  $C_1$  need not satisfy the Bragg condition exactly; its vertex must only be sufficiently close to the large circle shown in Fig. 3, say, within the region denoted by the smaller circle shown dashed.
- (2) In the above range of values for  $t_g/\lambda$ , both  $TE_0$  and  $TM_0$  modes exhibit strong blazing. This is expected because the propagation factors  $\beta$  for these modes are nearly equal, so that the Bragg construction in Fig. 4 holds for both modes.

Figure 5 also shows curves of the leakage parameter  $\alpha\lambda$ , which is proportional to the beam coupling factor. This parameter was found exactly by a numerical method (shown dashed) and by a simpler perturbation method,<sup>6</sup> (shown solid). It is interesting to note that leakage is larger in the range of  $t_g/\lambda$  for which blazing is maximum.

Dielectric gratings having other parameters are currently under investigation in order to further examine the generality of the present results.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

K.C. Chang and T. Tamir

National Science Foundation  
ENG 74-23908

#### REFERENCES

1. S.T. Peng and T. Tamir, *Optics Commun.*, Vol. 11, p. 405 (1974).
2. W. Streifer, D.R. Scifres and R.D. Burnham, *IEEE J. Quantum Electron.*, Vol. QE-12, p. 494 (1976).
3. D. Marcuse, *Bell Syst. Tech. J.*, Vol. 55, p. 1295 (1976).
4. T. Aoyagi, Y. Aoyagi and S. Namba, *Appl. Phys. Lett.*, Vol. 29, p. 303 (1976).
5. T. Tamir and S.T. Peng, "Network Methods for Integrated Optics Devices," *Proc. Intl. Conf. Holography and Data Processing* (Oxford: Pergamon Press, 1977), p. 437.
6. T. Tamir and S.T. Peng, *Appl. Phys.*, Vol. 14, p. 235 (1977).
7. H. Kogelnik and T.P. Sosnowsky, *Bell Syst. Tech. J.*, Vol. 49, p. 1602 (1970).

## EFFECTS OF PHASE VARIATIONS ON OPTICAL PARAMETRIC MODE-COUPLING DYNAMICS

E.S. Cassedy and M. Jain

Parametric mode-coupling phenomena in laser-driven plasmas has been studied for: (anomalous) heating mechanisms,<sup>1</sup> instability saturation mechanisms,<sup>2</sup> x-type wave breaking criteria<sup>3</sup> and interpretations of soliton formation.<sup>4</sup> Similar mode-coupling has also been considered for ionospheric applications,<sup>2,5</sup> for magnetically confined plasmas<sup>6</sup> and in the study of ocean wave dynamics.<sup>7</sup> In any of these studies, the possibility of multi-mode parametric<sup>†</sup> interactions must be considered, albeit the fastest-growing interaction only is often chosen as a simplifying approximation.

In the plasma literature the time-dynamics of parametric mode-coupling have customarily been treated as an initial-value problem. Such a treatment means typically<sup>1-3</sup> that the (large-amplitude) pump mode is assumed to be "turned on" (at  $t = 0$ ) and that the other (initially-small) interacting modes grow up to the saturation level thereafter. When multi-mode interactions are treated in this manner, the various modes have been assumed to belong to the Fourier-mode spectrum<sup>2-5</sup> of an unbounded plasma.

Pump-driven, bounded plasmas, on the other hand, have been considered for three-wave interactions<sup>8,9</sup> of the parametric type. Here the time-space evolution of the three (coupled) waves is assumed to be governed by three, coupled, first-order wave equations (in time and space) with appropriate boundary and initial values. Here the initial condition can assume zero-level pump fields and small amplitudes for the other two waves throughout the interior bounded region of interaction. The boundary values, however, include a fixed value for the pump field at one boundary to represent the pump wave incident from outside of the region. Thus the wave solutions in these formulations are, in effect, driven (or sustained) by the incident pump power.

The time dynamics of the parametric interaction of modes have been examined in the quantum electronics literature<sup>10-12</sup> for application to optical parametric oscillators (OPOs). The modes for an OPO are the resonant longitudinal modes of a Fabry-Perot cavity. First-order modelling equations for the time dynamics of OPOs have been given by Yariv and Louiselle,<sup>11</sup> who assume only a single-longitudinal-mode, three-frequency interaction. These equations have been generalized<sup>12</sup> for multi-mode, multi-frequency parametric interactions. Whereas this previous work had "injection tuning," of OPOs as its principal objective, the modelling equations appeared capable of dealing with multi-mode effects more generally, within the assumptions made.

---

<sup>†</sup>The term "parametric" is used here in the generic sense and includes parametric-decay, stimulated-scattering and purely-growing interactions.

It should be emphasized that these modelling equations represent the coupled modes of the OPO as they are driven by an incident pump field. The equations<sup>11,12</sup> are a coupled set first-order differential equations with an inhomogeneous (pump, driving) term, which is distinct from the pump mode amplitude (a dependent variable). Markedly different saturation occurs for the driven case as versus the undriven, as must be expected. In any driven multi-mode case, there is always a final steady state (limited by mode damping) where the fastest growing mode is dominant along with its idler mode and the pump mode. All other modes (and idlers) decay once the pump mode amplitude has been depleted to the threshold value. In the undriven case,<sup>†</sup> on the other hand, saturation seems to be dominated by the lightest-damped modes and not by the highest-gain modes.

The modelling equations<sup>11,12</sup> are, of course, deterministic in the complex amplitudes. As such, they would appear to ignore the effects of random phases<sup>13,14</sup> as versus well-defined phases,<sup>15</sup> which have resulted in significantly different predictions on stabilization of plasma wave-wave interactions. It should be noted here, however, that these previous studies<sup>13-15</sup> were for undriven systems. Our purpose here is to demonstrate that a driven set of time-dynamic equations, representing a bounded, parametric-wave system, has a semi-well-defined final steady state in the phases of the interacting modes. The phases are found to be "semi-well-defined," since only the pump-mode phase is uniquely determined, for the usual initial conditions (all modes grow from noise levels). The signal and idler modes, on the other hand, appear to meet a condition concerning the sum of their phases, but not have individual unique values in the final steady state.

In addition, the results to follow will demonstrate that the (mode) amplitude growth to saturation, for the entire range of possible initial (mode) phases, is bounded by a relatively narrow envelope in time. Furthermore, the (mode) amplitudes settle to unique values in the final (saturated) state, regardless of (pump, signal or idler) amplitudes. The results would suggest that the assumption of initial (mode) phases is unimportant for any consideration dealing solely with amplitude (or intensity) growth, to saturation, of a driven parametric system.

In our previous study of OPO dynamics<sup>12</sup> only absolute values of the signal, idler and pump mode amplitudes were plotted. Since the mode amplitudes are complex quantities, the phase dependence should be included in the dynamic behavior of the OPO. In order to study the effect of those phase variations, consider the simplest (single mode,

---

<sup>†</sup> Large initial amplitudes of the pump mode are assumed.

$N=1$ ) set of normalized (complex) equations:<sup>12</sup>

$$\frac{da_p}{dT} = -a_p - \sqrt{\alpha_1 \alpha_2} a_1 a_2 + i\beta \quad (1)$$

$$\frac{da_1}{dT} = -\alpha_1 a_1 + \sqrt{\alpha_1 \alpha_2} a_2^* a_p \quad (2)$$

$$\frac{da_2^*}{dT} = -\alpha_2 a_2^* + \sqrt{\alpha_1 \alpha_2} a_1 a_p^* \quad (3)$$

Separating the phases by writing,  $a_p = A_p e^{i\phi_p}$ ,  $a_1 = A_1 e^{i\phi_1}$  and  $a_2 = A_2 e^{i\phi_2}$ , one obtains a set of six real equations:

$$\frac{dA_p}{dT} = -A_p - \sqrt{\alpha_1 \alpha_2} A_1 A_2 \cos(\phi_1 + \phi_2 - \phi_p) + \beta \sin \phi_p \quad (4a)$$

$$\frac{dA_1}{dT} = -\alpha_1 A_1 + \sqrt{\alpha_1 \alpha_2} A_2 A_p \cos(\phi_1 + \phi_2 - \phi_p) \quad (4b)$$

$$\frac{dA_2}{dT} = -\alpha_2 A_2 + \sqrt{\alpha_1 \alpha_2} A_1 A_p \cos(\phi_1 + \phi_2 - \phi_p) \quad (4c)$$

$$A_p \frac{d\phi_p}{dT} = -\sqrt{\alpha_1 \alpha_2} A_1 A_2 \sin(\phi_1 + \phi_2 - \phi_p) + \beta \cos \phi_p \quad (4d)$$

$$A_1 \frac{d\phi_1}{dT} = -\sqrt{\alpha_1 \alpha_2} A_2 A_p \sin(\phi_1 + \phi_2 - \phi_p) \quad (4e)$$

$$A_2 \frac{d\phi_2}{dT} = -\sqrt{\alpha_1 \alpha_2} A_1 A_p \sin(\phi_1 + \phi_2 - \phi_p) \quad (4f)$$

The final steady state of these equations is obtained by setting,

$$\frac{d}{dT} (A_p, A_1, A_2, \phi_p, \phi_1, \phi_2) = 0$$

Putting this condition in Eqs. (4d) to (4f) gives,

$$\sin(\phi_1 + \phi_2 - \phi_p) = 0$$

$$\cos \phi_p = 0$$

with the solution,  $\phi_p = \pm \pi/2$  and  $\phi_1 + \phi_2 = \pm \pi/2, \mp \pi/2$ . To find the steady state ( $d/dT = 0$ ) condition for Eqs. (4a) to (4c) we then examine:

$$A_p = -\sqrt{\alpha_1 \alpha_2} A_1 A_2 \cos(\phi_1 + \phi_2 - \phi_p) + \beta \sin \phi_p \quad (5a)$$

$$\alpha_1 A_1 = \sqrt{\alpha_1 \alpha_2} A_2 A_p \cos(\phi_1 + \phi_2 - \phi_p) \quad (5b)$$

$$\alpha_2 A_2 = \sqrt{\alpha_1 \alpha_2} A_1 A_p \cos(\phi_1 + \phi_2 - \phi_p) \quad (5c)$$

Remembering that  $A_p, A_1, A_2 > 0$ , we conclude that  $\cos(\phi_1 + \phi_2 - \phi_p) = 1$  (the -1 value being nonphysical). Also  $\sin \phi_p = 1$  (the -1 value also being nonphysical). Therefore there is a unique solution for  $\phi_p$  and  $\phi_1 + \phi_2$ , namely:  $\phi_p = \pi/2$  and  $\phi_1 + \phi_2 = \pi/2$ .

On adding Eqs. (4e) and (4f) it is seen that the complete set of six equations can be represented now by five equations which depend only on  $(\phi_1 + \phi_2)$  and not on  $\phi_1$  and  $\phi_2$  independently. In general, in the final steady state, the values of  $\phi_1$  and  $\phi_2$  will depend on the initial values of these parameters and may be obtained by numerical solutions (see Figure 1). These values, however, will always be such that  $\phi_1 + \phi_2 = \pi/2$  in the final steady state.

On putting the values  $\phi_p = \pi/2$  and  $\phi_1 + \phi_2 = \pi/2$  in Eqs. (5a) to (5c) and solving simultaneously, we find:

$$\alpha_1 \alpha_2 A_1 A_2 = \alpha_1 \alpha_2 A_1 A_2 A_p^2 \quad \text{or} \quad A_p = 1$$

$$1 = -\sqrt{\alpha_1 \alpha_2} A_1 \frac{\sqrt{\alpha_1 \alpha_2}}{\alpha_2} A_1 + \beta = -\alpha_1 A_1^2 + \beta$$

or

$$A_1 = \left( \frac{\beta - 1}{\alpha_1} \right)^{1/2};$$

similarly

$$A_2 = \left( \frac{\beta - 1}{\alpha_2} \right)^{1/2}$$

Therefore we have found that the final solution is unique as far as the absolute values are concerned.

#### A. Intermediate Dynamics

Since the original six equations (4a) to (4f) are cross coupled, the intermediate values of  $A_p, A_1$  and  $A_2$  are expected to depend on the initial phases  $\phi_p, \phi_1$  and  $\phi_2$  chosen. However, as seen before, these equations reduce to an equivalent set of five equations containing  $(\phi_1 + \phi_2)$  and not  $\phi_1$  and  $\phi_2$  separately. Hence the intermediate dynamics are expected also to depend only on the initial values  $\phi_p$  and  $(\phi_1 + \phi_2)$ .

Let us first consider the variation of  $\phi_p$  under the assumption that  $A_p, A_1$  and  $A_2$  start initially from very small values. Near time  $T=0$ , all  $A_p, A_1$  and  $A_2$  are of the same order, and are still extremely small, and hence one can drop the  $\sqrt{\alpha_1 \alpha_2} A_1 A_2$  term compared to the  $A_p$  term. Hence at short times, when  $A_1$  and  $A_2$  are very small,

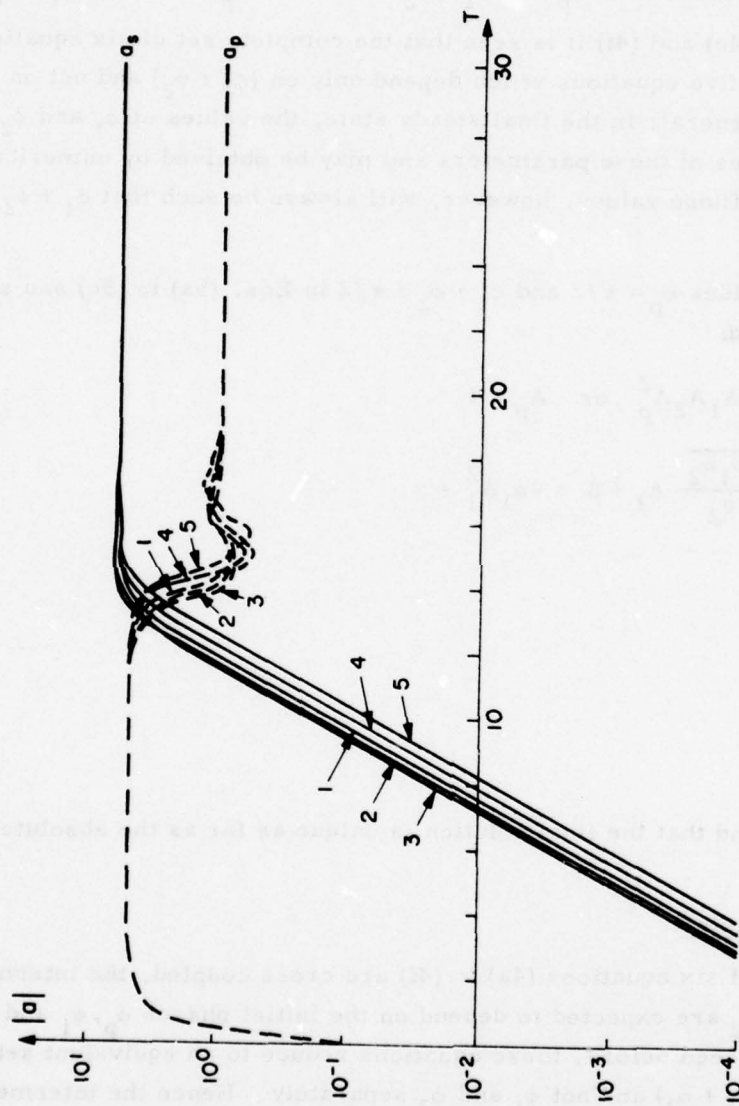


Fig. 1. Mode amplitudes versus time for parameters of initial phase (as tabulated in Table I).  
 $a_s$  = signal-mode amplitude (normalized<sup>1,2</sup>);  $a_p$  = pump-mode amplitude (normalized<sup>1,2</sup>)  
 $T$  = normalized<sup>1,2</sup> time.

TABLE I. Initial and final phases.

Fig. 1 curve number	initial values			final values			
	$\phi_p$	$\phi_1$	$\phi_2$	$\phi_p$	$\phi_1$	$\phi_2$	$\phi_1 + \phi_2$
①	0°	180°	0°	90°	154°	-64°	90°
	0°	0°	0°	90°	26°	64°	90°
②	0°	135°	0°	90°	121°	-31°	90°
	0°	45°	0°	90°	59°	31°	90°
③	0°	90°	0°	90°	90°	0°	90°
④	0°	-45°	0°	90°	-18°	108°	90°
	0°	-135°	0°	90°	-162°	252°	90°
⑤	0°	-90°	0°	90°	-90°	180°	90°

by the set of equations,

$$\frac{dA_p}{dT} = -A_p + \beta \sin \phi_p \quad (7a)$$

$$\frac{d\phi_p}{dT} = \frac{\beta}{A_p} \cos \phi_p \quad (7b)$$

with the steady state solution,  $\phi_p = \pi/2$ ,  $A_p = \beta$  (see Figure 1). Thus the pump reaches this quasi steady state shortly (a few normalized time constants) after time  $T=0$  and hence phase variation of  $\phi_p$  is not expected to take a significant part in changing these early-time dynamics. Later, when  $A_1$  and  $A_2$  become significant, one might expect  $\phi_p$  to vary from its steady state value ( $\pi/2$ ), as seen from Equation (4d). However, such variation would be small since the two terms on the RHS of Eq. (4d) are of opposite signs and once  $(\phi_1 + \phi_2) \rightarrow \pi/2$ ,  $d\phi_p/dT \rightarrow 0$ , even for large  $A_1$  and  $A_2$ . Thus variation in initial value of  $\phi_p$  is expected to play little role in changing the intermediate or final steady state dynamics.

However, the intermediate state does depend on the initial value of  $\phi_1 + \phi_2$  and numerical methods may be employed in order to obtain the actual values of  $a_1$  and  $a_2$ . Even before obtaining numerical results, one can make some qualitative deductions on examining the Equations (4b) and (4c). Since the parametric amplification term is

proportional to  $\cos(\phi_1 + \phi_2 - \phi_p)$ , one expects that the closer the initial  $(\phi_1 + \phi_2)$  is to the final steady state value  $(\pi/2)$ , since  $\phi_p = (\pi/2)$ , the more will be the amplification. In other words, the initial  $\phi_1 + \phi_2 = \pi/2$  curve for  $A_1$  will lie above the other curves; with the initial  $\phi_1 + \phi_2 = \pi/2$  curve being at the bottom. Such behavior is in fact confirmed by the numerical solutions shown on Fig. 1, for the initial phases listed in Table I.

Also, on changing  $\phi_1 + \phi_2 \rightarrow \pi - (\phi_1 + \phi_2)$ , the equations remain unchanged. Thus, any initial  $(\phi_1 + \phi_2)$  and  $\pi - (\phi_1 + \phi_2)$  will result in the same dynamical behavior. This is also shown by the numerical result. Figure 1 shows the dynamical behavior for  $(\phi_1 + \phi_2)$  initial value varying from  $\pi \rightarrow -3\pi/4$ . It is found (Table I) that the range  $\pi/2 \rightarrow -\pi/2$  covers the entire set of values as far as the values  $A_p, A_1$  and  $A_2$  are concerned, and the curves lie within a narrow bandwidth of each other.

It is to be noted that on changing  $i\beta \rightarrow \beta$ , one obtains an analogous set of equations. These equations have similar steady state solutions with  $a_p = 1$ ,  $A_1 = (\beta - 1/\alpha_1)^{1/2}$  and  $A_2 = (\beta - 1/\alpha_2)^{1/2}$ . The only differences is in the steady state values of the phases, which are:  $\phi_p = 0$ ,  $\phi_1 + \phi_2 = 0$ , respectively.

Los Alamos Scientific Laboratory  
Purchase Order NP6-21964-1

E.S. Cassedy

Joint Services Technical Advisory Committee  
F44620-74-C-0056

#### REFERENCES

1. W.L. Kruer, et al., "Nonlinear Behavior of Light-Driven Plasma Instabilities," Rensselaer Polytechnic Laser Interaction Conf. (1973) (also, Livermore Lab. preprint UCRL-74947).
2. J.J. Thompson, et al., "Mode-Coupling Saturation of the Parametric Instability and Electron Heating," Phys. Rev. Lett., Vol. 31, pp. 918-920 (1973).
3. D.W. Forslund, et al., "Theory of Stimulated Scattering Processes in Laser-Irradiated Plasmas," Phys. Fluids, Vol. 18, pp. 1002-1030 (1975).
4. W.M. Manheimer and K. Papadopoulos, "Interpretation of Soliton Formation and Parametric Instabilities," Phys. Fluids, Vol. 18, pp. 1397-1398 (1975).
5. D.W. Forslund, et al., "Parametric Excitation of Electromagnetic Waves," Phys. Rev. Lett., Vol. 29, pp. 249-252 (1972).
6. M. Porkolab, "High-Frequency Parametric Wave Phenomena and Plasma Heating: a Review," Physica, Vol. 82C, pp. 86-110 (1976).
7. B.J. West, et al., "Mode Coupling Description of Ocean Wave Dynamics," Phys. Fluids, Vol. 17, pp. 1059-1067 (1974).
8. W.M. Manheimer, "Connection Between the Linear and Nonlinear Theory of Anomalous Reflection from a Plasma Slab," Phys. Fluids, Vol. 17, pp. 1634-1635 (1974).
9. R.W. Harvey and G. Schmidt, "Three Wave Backscatter in a Finite Region," Phys. Fluids, Vol. 18, pp. 1395-1396 (1975).

10. L. B. Kruezer, "Single-Mode Oscillation of a Pulsed Singly Resonant Optical Parametric Oscillator," *Appl. Phys. Lett.*, Vol. 15, pp. 263-265 (1969); also, S. E. Harris (a review article) *Proc. IEEE*, Vol. 57, pp. 2096-2113 (1969).
11. A. Yariv and W. H. Louiselle, "Theory of the Optical Parametric Oscillator," *IEEE J. Quant. Electron.*, Vol. QE-2, pp. 418-424 (1966).
12. E. S. Cassedy and M. Jain, "A Theoretical Study of Injection Tuning of Optical Parametric Oscillators," submitted for publication, *IEEE J. Quant. Electron.* (May 1978); also, Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77, pp. 205-211 (November 1977).
13. R. C. Davidson, "Methods in Nonlinear Plasma Theory," (New York: Academic Press, 1972).
14. R. E. Aamodt, et al., "Nonlinear Dynamics of Drift-Cyclotron Instability," *Phys. Rev. Lett.*, Vol. 39, pp. 1660-1664 (1977).
15. H. Wilhelmsson, et al., "Explosive Instabilities in the Well-Defined Phase Description," *J. Math. Phys.*, Vol. 11, pp. 1738-1742; F. Engelmann and H. Wilhelmsson, "Phase Effects in the Nonlinear Interaction of Negative-Energy Waves," *Zeit. fur Naturforsch.*

## THEORY OF THE INTEGRATING SPHERE FOR PULSED LIGHT SOURCES

K. Park and W. T. Walter

An integrating sphere is frequently used in the measurement of the diffuse reflectance of a surface because of its ability to collect light reflected into all angles by the target surface. A continuous light source is employed in these measurements so that time delays caused by multiple reflections within the integrating sphere affect only the initial detector response and not the steady state values. In examining the reflectance change of a surface during a short intense pulse of light such as a Q-switched laser pulse, however, the effect of multiple reflections must be taken into consideration.

Not only does an integrating sphere carry out a spatial integration of light reflected from a target surface, but it delays the arrival of light at the detector by varying times depending on the path within the sphere. Thus the temporal shape of the detected light signal may be significantly modified. For example, we have found<sup>1</sup> that a 30-ns FWHM (Full Width at Half-Maximum intensity) Q-switched ruby laser pulse reflected from a target surface placed at the sample port of a 20 cm diam integrating sphere was lengthened by  $\sim 10$  ns when measured by a fast vacuum photodiode located at the detector port (Figure 1).

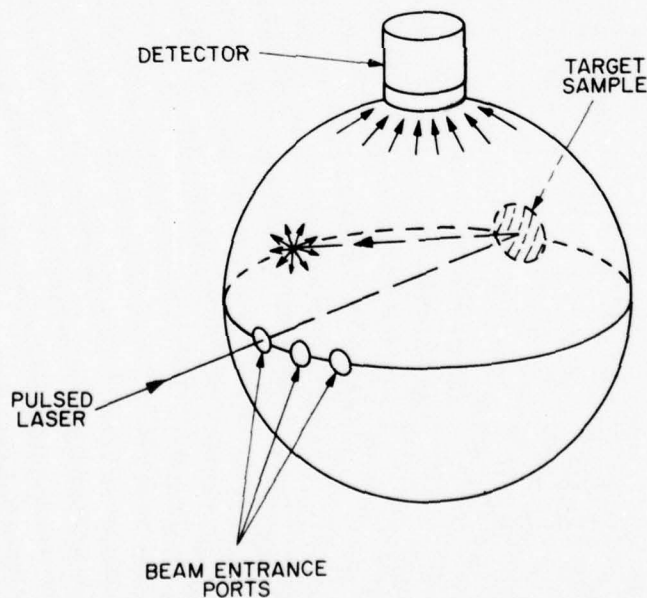


Fig. 1. Total reflectance measurement of a metal surface using an integrating sphere.

At first thought it might seem that if the same process were applied to a sample of the incident light signal, then the ratio of the two output signals from the integrating

sphere would still yield the reflectance of the target surface. It will be pointed out in this report that this is only true for slowly changing reflectances.

So far as the authors are aware, no theoretical analysis of the integrating sphere for a time-dependent light source is available. Jacquez and Kuppenheim<sup>2</sup> give the general theory of the integrating sphere for time-independent light sources using an integral equation formulation. They solved the equation using Fredholm's method. In this report their theory is extended to include time-dependent light sources, such as fast laser pulses.

The theory of the integrating sphere is the theory of multiple reflections in a cavity. Consider an integrating sphere with radius  $R$  whose interior wall is a perfect diffuse reflector (i.e., all points on the surface have constant radiance). A light pulse with power  $P_0(t)$  [watts] is directed into the cavity through an entrance window (Fig. 1) onto the wall where it begins a path of multiple reflections before collection by the detector. We shall assume that the total area of the various ports is small compared to the total area of the sphere and that the reflecting surface is perfectly spherical.

Using the spherical coordinates  $(\theta, \phi)$ , we shall represent by  $I(t, \theta, \phi)$  the irradiance at  $(\theta, \phi)$  and at time  $t$  caused by the reflected flux in the sphere. Following Jacquez and Kuppenheim<sup>2</sup> we shall treat the irradiance striking an area  $da$  on the sphere as a sum of two components,  $I_0(t, \theta, \phi)$  representing the primary incident light pulse and  $I(t, \theta, \phi)$  representing the light reflected by the remainder of the sphere to  $da$ . The quantity  $I_0(t, \theta, \phi)$  will be zero for all points on the wall of the sphere except for the small area directly irradiated by the incident light beam. Let a constant  $r$  denote the total reflectivity at any point on the sphere. Then  $r(I_0 + I)$  is reflected. Since all points on the surface are assumed to reflect perfectly diffusely, any element area of  $da$  (Fig. 2) has constant radiance  $N$  and

$$N = [I_0(t, \theta, \phi) + I(t, \theta, \phi)] \frac{r}{\pi} \quad (1)$$

Now the flux radiated from one area  $da'$  to a second area  $da$  is given by  $N' d\Omega da' \cos \alpha$  where  $d\Omega$  is the solid angle subtended by the area element  $da$  and  $\cos \alpha$  is the Lambertian factor (Figure 2). From Fig. 2 and Eq. (1) we can show that the radiated flux from  $da'$  to  $da$  is

$$\frac{r}{\pi} [I_0(t, \theta', \phi') + I(t, \theta', \phi')] da \frac{(\underline{\rho} \cdot \underline{n})}{\rho^3} da' \frac{(-\underline{\rho} \cdot \underline{n}')}{\rho} \quad (2)$$

On the sphere,

$$\frac{(\underline{\rho} \cdot \underline{n})}{\rho^3} \frac{(-\underline{\rho} \cdot \underline{n}')}{\rho} = \frac{1}{4R^2}$$

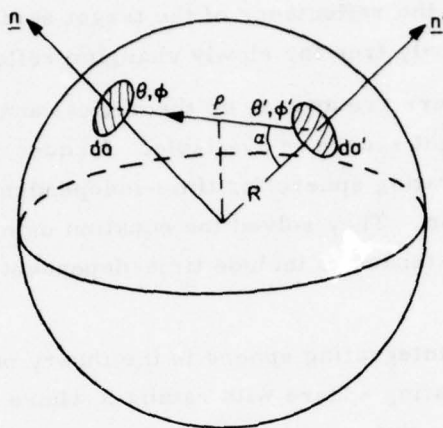


Fig. 2. Vector relations between area elements on the sphere.

Therefore, the irradiance at  $(t, \theta, \phi)$  is given by

$$I(t, \theta, \phi) = \frac{r}{4\pi R^2} \int_S I_0(t - \Delta t, \theta', \phi') da' + \frac{r}{4\pi R^2} \int_S I(t - \Delta t', \theta', \phi') da' \quad (3)$$

where  $\Delta t$ , is the time for light to travel from  $(\theta', \phi')$  to  $(\theta, \phi)$  and is given by

$$\Delta t = \frac{\rho}{c} = \frac{2R \cos \alpha}{c}$$

where  $c$  is the speed of light.

The first integral on the right-hand side of Eq. (3) gives  $r P_0(t)/4\pi R^2$  where  $P_0(t)$  is the incident light power. Here an approximation has been made that  $I_0(t - \Delta t, \theta', \phi') \sim I_0(t, \theta', \phi')$  which is valid for a laser pulse whose FWHM is much larger than  $R/c$ . For a 20 cm diam integrating sphere,  $\rho/c \leq 0.7$  ns.

Thus, we have a time-delayed integral equation:

$$I(t, \theta, \phi) = r I_s(t) + \frac{r}{4\pi R^2} \int_S I(t - \Delta t, \theta', \phi') da' \quad (4)$$

where

$$I_s(t) = \frac{P_0(t)}{4\pi R^2}$$

Since  $I_s(t)$  is not a function of position on the surface of the sphere, the only dependence of  $I(t, \theta, \phi)$  on position is through the time delay  $\Delta t(\theta, \phi, \theta', \phi')$ . If we take the observation point  $(\theta, \phi)$  at the north pole (any point is equivalent on the sphere for our assumptions), then the  $\phi'$  dependence disappears since  $\Delta t = \frac{2R}{c} \sin \frac{\theta'}{2}$  (see Figure 3).

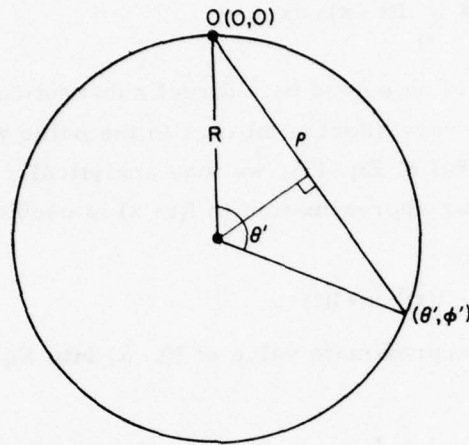


Fig. 3. Relations between the observation point and the source point.

Now substituting into Eq. (4) we find that

$$I(t) = r I_s(t) + \frac{r}{2} \int_0^\pi I(t - \sin \frac{\theta'}{2}) \sin \theta' d\theta' \quad (5)$$

where a new time unit of  $2R/c$  has been employed.

Changing the variable in Eq. (5) by putting  $\sin \frac{\theta'}{2} = x$ , then we obtain

$$\sin \theta' = 2 \sin \frac{\theta'}{2} \cos \frac{\theta'}{2} = 2x \sqrt{1-x^2}$$

$$d\theta' = \frac{2 dx}{\sqrt{1-x^2}}$$

and Eq. (5) becomes

$$I(t) = r I_s(t) + 2r \int_0^1 I(t-x) x dx \quad (6)$$

At this point we shall consider absorption and aperture losses of the integrating sphere. Let the total area of the sphere be  $S \equiv 4\pi R^2$ , the spherical area removed for the detector port be  $b$ , that for the entrance ports  $a$ , and that for the sample port  $c$ . Then, if  $d \equiv S - (a + b + c)$ , we may define an equivalent average sphere reflectance  $\ell \equiv rd/S + r_s c/S$  where  $r_s$  is the average reflectance of the sample attached to the sample port. The quantity  $(1 - \ell)$  accounts for total losses due to the absorption by the integrating sphere wall and target sample as well as the loss due to escape of light flux through the various holes. Thus, Eq. (6) can be written

$$I(t) = r I_s(t) + 2\ell \int_0^1 I(t-x) x dx . \quad (7)$$

Equation (7) could be solved by a direct substitution method.<sup>3</sup> However, since the integration interval is very short compared to the pulse width, instead of directly substituting into the integral of Eq. (7), we may analytically carry out the integration in the following way. A linear approximation to  $I(t-x)$  is used over the range from  $x=0$  to  $x=1$ . Then

$$I(t-x) = [I(t-1) - I(t)] x + I(t) .$$

Now, substituting the approximate value of  $I(t-x)$  into Eq. (7) and carrying out the integration, we obtain

$$(1 - \frac{1}{3}\ell) I(t) = r I_s(t) + \frac{2}{3}\ell I(t-1) . \quad (8)$$

Finally, we may express  $I(t)$  as

$$I(t) = c_1 I_s(t) + c_2 I(t-1) \quad (9)$$

where

$$c_1 = \frac{3r}{3-\ell} ,$$

$$c_2 = \frac{2\ell}{3-\ell} ,$$

$$I_s(t) = \frac{P_o(t)}{S} ,$$

and

$$\ell = \frac{rd}{S} + \frac{r_s c}{S} .$$

To obtain  $I(t)$  from Eq. (9) we follow an iterative procedure as follows:

$$I^{(0)}(t) = c_1 I_s(t)$$

$$I^{(1)}(t) = c_1 I_s(t) + c_2 I^{(0)}(t-1)$$

$$= c_1 I_s(t) + c_1 c_2 I_s(t-1)$$

$$I^{(2)}(t) = c_1 I_s(t) + c_2 I^{(1)}(t-1)$$

$$\begin{aligned}
&= c_1 I_s(t) + c_1 c_2 I_s(t-1) + c_1 c_2^2 I_s(t-2) \\
&\vdots \\
I^{(n)}(t) &= c_1 I_s(t) + c_2 I^{(n-1)}(t-1) \\
&= c_1 [I_s(t) + c_2 I_s(t-1) + c_2^2 I_s(t-2) + \dots + c_2^n I_s(t-n)]
\end{aligned} \tag{10}$$

This series converges for  $c_2 = \frac{2\ell}{3-\ell} < 1$ , i.e., for  $\ell < 1$ . Therefore, whenever there is a loss in the integrating sphere (which is always the case) the series converges. From the last expression of Eq. (10) we can clearly see how the incident pulse  $I_s(t)$  transforms itself into a longer pulse for each reflection inside the sphere.

Conversely, we shall now address the more important question of whether it is possible to recover the original pulse from the detected pulse which emerges from the integrating sphere. Equation (9) can be used to find the answer without having to trace back through the iterations indicated in Equation (10). From Eq. (9) then,

$$I_s(t) = \frac{1}{c_1} [I(t) - c_2 I(t-1)] \quad , \tag{11}$$

and therefore the inverse problem is actually the easier one. For pulses whose  $\text{FWHM} = \frac{2R}{c}$  or less, however, Eq. (7) must be used to find  $I_s(t)$ . Then,

$$I_s(t) = \frac{1}{r} [I(t) - 2\ell \int_0^1 I(t-x)x dx] \quad . \tag{12}$$

The output signal  $V(t)$  of a sufficiently-fast photodetector is given as  $V(t) = Db I(t)$  where  $D$  is the detector responsivity in volts/watt and  $b$  represents the area of the detector. Using  $I_s(t) = P_o(t)/(4\pi R^2)$  we obtain from Eq. (11):

$$P_o(t) = \frac{4\pi R^2}{D c_1 b} [V(t) - c_2 V(t - \frac{2R}{c})] \tag{13}$$

where  $P_o(t)$  is the original light flux entering the integrating sphere and we have returned to the conventional time unit of seconds.

For a 20 cm diam integrating sphere coated with a barium sulfate coating<sup>4</sup> having 99% reflectance at 700 nm and with holes for the detector, target sample and beam entrance port amounting to 3% of the total surface area of the sphere

$$\ell \sim .96, c_1 \sim 1.46, c_2 \sim .948 \quad .$$

Now from Eq. (13)

$$V_o(t) = 57[V(t) - .948 V(t - 0.67)] \quad (14)$$

where  $V_o(t)$  is the voltage that would appear on the photodetector output if there were no delay due to multiple reflections within the integrating sphere. In Eq. (14),  $t$  is measured in nanoseconds. Equation (14) can be used to recover the voltage  $V_o(t)$  of a light pulse entering the integrating sphere from the voltage  $V(t)$  displayed on the oscilloscope from the detector.

Summation of multiple reflections resulting in a final output pulse from the integrating sphere, according to the theory just developed, is shown in Fig. 4, where the

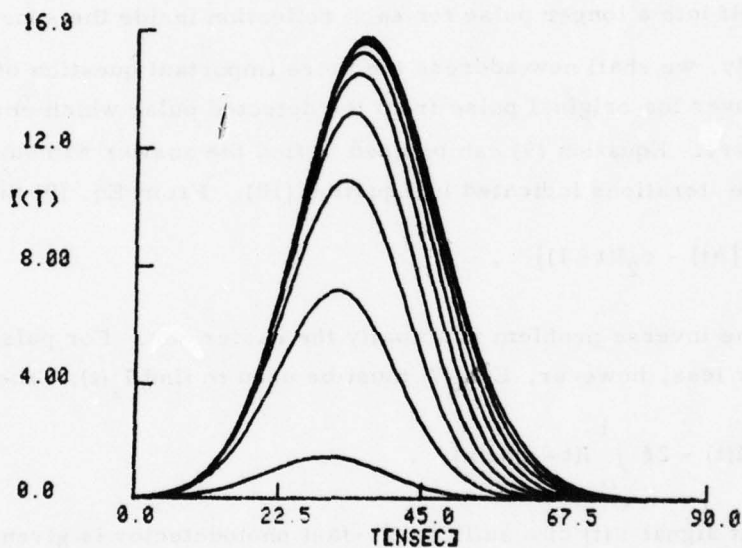


Fig. 4. Development of an output pulse from the integrating sphere.

development of an output pulse is shown by successive pulse curves. The smallest pulse curve is a portion of the original pulse intensity, that is,  $I_s(t)$ , the component directly reflected into the detector. Each successive curve shown represents five iterations (in Equation (10)). A total of 40 iterations carried out on a PDP 11/40 mini-computer are displayed in Figure 4. The loss  $(1 - \ell)$  was taken as 5%. After ~100 iterations the pulse shape has reached a steady state.

In Fig. 5(a) and (b), the development of the lengthened pulses exiting from the sphere and the recovery of the original pulses entering the sphere are illustrated for pulse curves which have been chosen to model the experimental results of Bonch-Bruevich<sup>6</sup> and Zavec<sup>7</sup>. They found that the reflectance of metal surfaces decreased

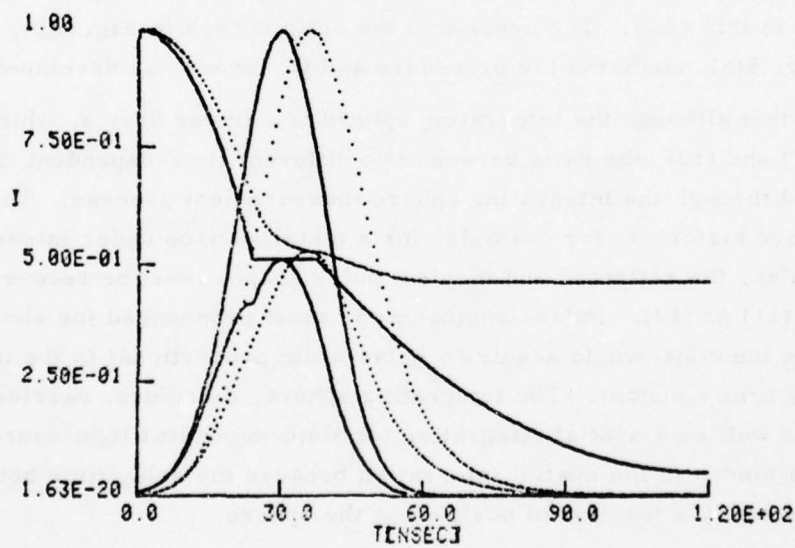
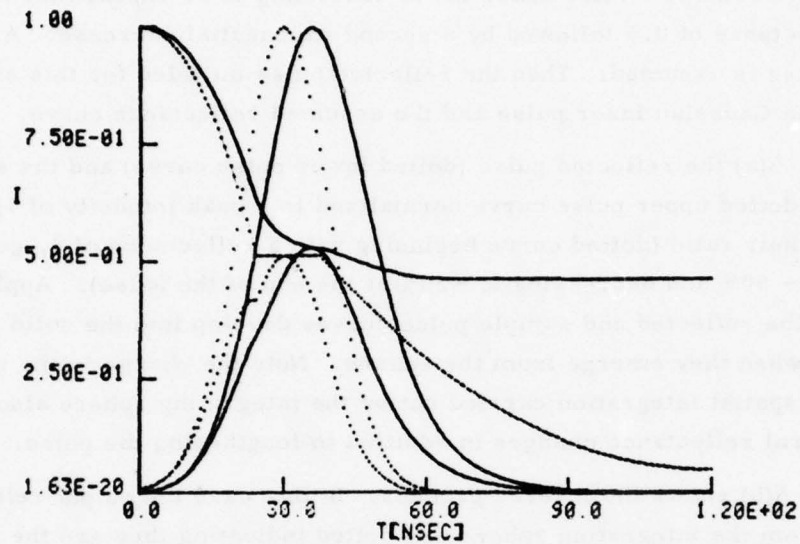


Fig. 5. (a) Calculation of sample pulse, reflected pulse and reflectance for pulses exiting the integrating sphere, (b) Recovery of sample pulse, reflected pulse and reflectance for pulses entering the integrating sphere.

dramatically during an intense laser pulse. To model the results of Bonch-Bruевич and Zavec we assume a reflectance curve consisting of an exponential decrease to a plateau reflectance of 0.5 followed by a second exponential decrease. A Gaussian input laser pulse is assumed. Then the reflected pulse modeled for this situation is the product of the Gaussian laser pulse and the assumed reflectance curve.

In Fig. 5(a) the reflected pulse (dotted lower pulse curve) and the sample incident laser pulse (dotted upper pulse curve normalized to a peak intensity of 1) are shown together with their ratio (dotted curve beginning with a reflectance of 1, going through a plateau at  $R \sim 50\%$  and decreasing to  $\sim 20\%$  at the end of the pulse). Applying Eq. (10), we find that the reflected and sample pulse curves develop into the solid curves shown in Fig. 5(a) when they emerge from the sphere. Note the change in the reflectance curve. The spatial integration carried out by the integrating sphere also smooths the sharp temporal reflectance changes in addition to lengthening the pulse.

Figure 5(b) shows the inverse process. In this case the output reflected and sample pulses from the integrating sphere are dotted indicating they are the initial conditions. Recovery of the input reflected and sample pulses entering the integrating sphere is obtained using Eq. (11) and is indicated by the solid curves in Figure 5(b). No iteration is necessary in this case. The overlap of the solid curves in Fig. 5(b), with the dotted curves of Fig. 5(a), confirms the procedure and its inverse as developed in this report.

Notice that although the integrating sphere is a linear device, which is apparent from Eqs. (7) and (10), the ratio between two different time-dependent light sources is not preserved through the integrating sphere measurement process. Therefore to obtain reflectance histories, for example, for a metal surface under intense irradiation by a laser pulse, the reflected and incident pulse shapes must be recovered according to Equations (11) or (12). Pulse-lengthening is most pronounced for short pulses. A delta-function input  $\delta(t)$  would acquire a pulse width proportional to the integrating sphere cavity time constant. The integrating sphere, therefore, carries out a time integration as well as a spatial integration for time-dependent light sources. The time integration is hidden in the spatial integration because the delay time between various reflections is really a function of position on the sphere.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

Air Force Office of Scientific Research  
AFOSR-78-3557

K. Park and W. T. Walter

## REFERENCES

1. K. Park, "Change in Reflectivity of a Metal Surface During Intense Laser Irradiation," Ph.D. Thesis, Polytech. Inst. of New York (1978).
2. John A. Jacquez and Hans F. Kuppenheim, "Theory of the Integrating Sphere," J. Opt. Soc. Am. 45, 460-470 (1955).
3. W. V. Lovitt, "Linear Integral Equations," Dover Publications, Inc., New York (1950).
4. Eastman White Reflectance Coating, Catalog No. 6030, Eastman Kodak Co., Rochester, New York 14650. In order to have an adequate thickness of the coating, the inside surfaces of the two hemispheres were sprayed four to eight times using a Kodak Laboratory Sprayer, Catalog No. 13270.
5. Eastman White Reflectance Coating, Kodak Publication No. JJ-32, Rochester, New York (1976); and F. Grum and G. W. Luckey, "Optical Sphere Paint and a Working Standard of Reflectance," Appl. Optics 7, 2289-2294 (1965).
6. A. M. Bonch-Bruевич, Ya. A. Imas, G. S. Romanov, M. N. Libenson and L. N. Mal'tsev, "Effect of a Laser Pulse on the Reflecting Power of a Metal," Sov. Phys-Tech. Phys. 13, 640-643 (1968).
7. T. E. Zavecz, M. A. Saifi and M. Notis, "Metal Reflectivity under High-Intensity Optical Radiation," Appl. Phys. Lett. 26, 165-168 (1975).

## TWO-PHOTON CORRELATIONS IN AMPLIFIED LASER LIGHT

G. D. Blake and D. B. Scarl

We have measured the two-photon correlation function in 632.8 nm single mode He-Ne laser light which had been amplified by stimulated emission in a long He-Ne discharge tube. For laser light intensities which were comparable with the discharge tube spontaneous emission intensity into a single mode, the measured correlations agree well with those predicted for a simple mixture of laser and chaotic light. The results are consistent with the assumption that laser light, linearly amplified by stimulated emission, has the same statistical properties as unamplified laser light.

A. Introduction

Although the statistical properties of simple electromagnetic fields are well understood, the interpretation of field measurements in terms of elementary physical concepts is not always straightforward. One of several sources of this difficulty is that the mode operators into which the field operators most easily separate have eigenfunctions whose squares are independent of position and time. Using these mode operators, the usual interpretation of the field in terms of photons yields photons that are eigenstates of the Hamiltonian and therefore independent of time.

Photons that are excitations of a single field oscillator at a single frequency necessarily extend over all time, a requirement that is difficult to reconcile with the creation or destruction of these photons. On the other hand, the photons that participate in experiments seem to be created in the neighborhood of one space time point, to propagate occupying a finite region of space-time, and to be detected in the neighborhood of another space-time point. It is these photons which carry information, excite atoms, show interference effects, and display photon statistics.

The properties of photons are intimately connected with their interaction with atoms and other massive objects. The fundamental processes of absorption, spontaneous emission, and stimulated emission seem to exist in all sources, converters, filters, and detectors. The dynamics of these processes and their connection with the properties of the absorbed or emitted fields has been one of the most beautiful applications of quantum mechanics in the past two decades. However, the time dependence of these transitions is usually described without the use of temporally localized photons.

Stimulated emission illustrates one of the apparent difficulties of current photon descriptions of the electromagnetic field. When an identifiable incident photon stimulates an emitted photon, the incident and emitted photons might be expected to be related in time. Previous attempts to detect these time correlations in amplified

chaotic light have shown no correlations in excess of the HBT effect.<sup>1</sup> Since the HBT effect shows the same ratio of correlated photon pairs to uncorrelated photon pairs, independent of the geometry of the source or the light intensity which the atoms see, it is difficult to attribute the HBT effect itself entirely to stimulated emission. It can always be attributed to random correlations between photon pairs,<sup>2, 3</sup> although this point is still under discussion.<sup>4-8</sup>

In particular, single mode laser light, which can be described by a coherent state of the electromagnetic field, shows no HBT effect. Linearly amplified laser light, which can still be described by a coherent state, although having been amplified by stimulated emission, would therefore be expected to show no HBT effect and no time correlations between photon pairs.

We have passed a beam from a single mode laser through an amplifier tube in which stimulated emission increases the intensity of the laser beam by 30%, and have looked for time correlations between pairs of successive detected photons in the output light. The amplified laser beam was of approximately the same intensity as the spontaneously emitted light from the gain tube in the same spaceangle mode. Under these conditions every stimulated photon generated by the original laser beam is accounted for in the amplified laser light; there is no other stimulated emission associated with the laser light. The observed photon correlations can be attributed to second order interference between pairs, one of which is from the amplified laser light, and one of which is from amplified spontaneous emission, or both of which are from amplified spontaneous emission. No correlations between incident and induced photons seem to be necessary in order to explain the results.

#### B. Counting Rate for Point Detectors

For a beam with stationary statistics, the two-photon counting rate for two infinitesimal detectors, one of which samples the field at the space-time point  $x_1 = (\vec{r}_1, t_1)$ , and the other of which samples the field at the point  $x_2$ , is proportional<sup>10</sup> to the second order correlation function:

$$\begin{aligned} G^{(2)}(x_1, x_2) &= \langle A^\dagger(x_1) A^\dagger(x_2) A(x_1) A(x_2) \rangle \\ &= \text{Tr} \rho A(x_1) A^\dagger(x_2) A(x_1) A(x_2) \end{aligned} \quad (1)$$

where  $A(x)$  is the field operator at the point  $x$  and  $\rho$  is the field density operator. The field operators can be expanded in a set of discrete orthogonal modes,  $q(k_i, x)$  with the mode amplitude operators  $a_i$  and propagation constants  $k_i = (\vec{k}_i, \omega_i)$ .

$$A(x) = \sum_i a_i q(k_i, x) \quad (2)$$

The field density operator in the present experiment in which the amplified laser light has approximately the same intensity as spontaneous emission from the amplifier tube, describes a mixture of amplified laser and chaotic (gain tube spontaneous emission) light. The amplified laser light is assumed to have the same density operator as ordinary unamplified single mode laser light. The chaotic light has the usual maximum-entropy density matrix. The density matrix for the mixture, written in a basis made up of the coherent states for each mode,  $|a_l\rangle$  is<sup>11</sup>

$$\rho = \pi \int P_l(a_l) |a_l\rangle \langle a_l| d^2 a_l \quad (3)$$

It is convenient to let the mode  $l = 1$  be that which contains the laser light. For this mode

$$P(a_1) = \frac{1}{\pi \langle n_1 \rangle} e^{-|a_1 - a_0|^2 / \langle n_1 \rangle} \quad (4)$$

where  $|a_0|^2$  is the average number of photons in the amplified laser beam and  $\langle n_1 \rangle$  is the number of chaotic photons in the amplified spontaneous emission mode with the same central frequency and propagation vector as the laser light.

For all other modes there is spontaneous emission alone and the weight function is

$$P_l(a_l) = \frac{1}{\pi \langle n_l \rangle} e^{-|a_l|^2 / \langle n_l \rangle} \quad (5)$$

where  $\langle n_l \rangle$  is the average number of spontaneously emitted photons with frequency  $\omega_l$  and propagation vector  $\vec{k}_l$ .

The second order correlation function can be written

$$\begin{aligned} G^{(2)}(x_1, x_2) &= \text{Tr} \pi \int P_l(a_l) |a_l\rangle \langle a_l| \sum_{i,j,k,m} a_i^\dagger a_j^\dagger a_k a_m q^*(k_i, r_1) q^*(k_j, r_2) q(k_k, x_1) q(k_m, x_2) d^2 a_l \\ &= \pi \sum_{i,j,k,m} \int a_i^* a_j^* a_k a_m P_l(a_l) q^*(k_i, x_1) q^*(k_j, x_2) q(k_k, x_1) q(k_m, x_2) d^2 a_l \end{aligned} \quad (6)$$

Equation (6) is also the classical expression for the second order correlation function of a field with mode amplitudes  $a$  whose distribution functions are  $P(a)$ .

Using the fact that integrals containing unpaired  $a$ 's are zero, the sums and products of Eq. (6) reduce to

$$\begin{aligned} G^{(2)}(x_1, x_2) = & \sum_i \int |a_i|^4 P_i(a_i) d^2 a_i + \\ & \left| \sum_i \int |a_i|^2 P_i(a_i) d^2 a_i \right|^2 + \\ & \left| \sum_i \int |a_i|^2 P_i(a_i) d^2 a_i q^*(k_i, x_1) q(k_i, x_2) \right|^2. \end{aligned} \quad (7)$$

After considerable rearrangement, reconstruction of sums, and replacement of sums by integrals,<sup>12</sup>

$$G^{(2)}(x_1, x_2) = (I_L + I_C)^2 + I_C^2 |G^{(1)}(x_1, x_2)|^2 + 2I_L I_C \operatorname{Re} G^{(1)}(x_1, x_2) \quad (8)$$

where

$$I_C = \frac{1}{k} \int \langle n_k \rangle dk \quad (9)$$

$$I_L = |a_0|^2 \quad (10)$$

$$G^{(1)}(x_2, x_1) = \frac{1}{k} \int \langle n_k \rangle e^{ik(x_2 - x_1)} dk. \quad (11)$$

$I_L$  and  $I_C$  are the laser and chaotic light intensities.  $G^{(1)}(x_1, x_2)$  is the first order correlation function for the field. For a stationary, cross-spectrally pure field, the correlation functions depend only on coordinate differences and the first order correlation function factors into separate functions of time and of space:

$$\begin{aligned} G^{(2)}(\Delta r, \Delta t) = & (I_L + I_C)^2 \left[ 1 + \frac{I_C^2}{(I_L + I_C)^2} |G^{(1)}(\Delta r)|^2 |G^{(1)}(\Delta t)|^2 + \right. \\ & \left. 2 \frac{I_L I_C}{(I_L + I_C)^2} \operatorname{Re} G^{(1)}(\Delta r) G^{(1)}(\Delta t) \right]. \end{aligned} \quad (12)$$

The two-photon counting rate can be interpreted as a background term, independent of the detector spacing, a term arising from correlations between pairs of chaotic photons, and a term arising from correlations between chaotic photons and photons in the amplified laser beam.

### C. Expected Counting Rate for Real Detectors

The expected counting rate for detectors which have a finite spatial extent and a finite time resolution can be gotten by integrating over all pairs of points in the detector aperture and convolving with the detector two-photon time resolution function.

After integrating over the second aperture, convolving with the photomultiplier time response and taking account of noise, filter leakage, and detector efficiency, the expected counting rate is

$$R = R_o \left[ 1 + C_1 e^{-\Delta t^2 / 2\sigma_1^2} + C_2 e^{-\Delta t^2 / 2\sigma_2^2} \right] \quad (13)$$

where

$$R_o = I^2 \epsilon_1 \epsilon_2 T \quad (14)$$

$$I = I_L + I_C + I_N + I_W \quad (15)$$

$$C_1 = S_1 \frac{\sigma_t / \sqrt{2}}{\sigma_1} \frac{I_C^2}{I^2} \quad (16)$$

$$C_2 = S_2 \frac{\sigma_t}{\sigma_2} \frac{I_L I_C}{I^2} \quad (17)$$

Here  $\epsilon_1$  and  $\epsilon_2$  are the detection efficiencies of the two detector chains,  $T$  is the width of one PHA time bin,  $I_L$ ,  $I_C$ , and  $I_W$  are the number of photons per second originating from the laser, from chaotic 632.8 nm light from the amplifier, and from chaotic unwanted 633.4 nm light from the amplifier.  $I_N$  is the mean number of photomultiplier noise counts divided by  $\epsilon_1$ . The correlation terms,  $C_1$  and  $C_2$  are the correlated events (as a fraction of  $R_o$ ) that arise from pairs of chaotic 632.8 nm photons and from 632.8 nm chaotic and 632.8 nm laser photons.

The spatial coherence factors  $S_1$  and  $S_2$  arise from integrals of  $G^{(1)}(\Delta r)$  and  $|G^{(1)}(\Delta r)|^2$  over the detector aperture. They are evaluated and plotted in References 13-15.  $\sigma_t$  is the standard deviation of the Gaussian first order time correlation function for the chaotic light, and

$$\sigma_1 = \sqrt{(\sigma_t / \sqrt{2})^2 + \sigma^2} \quad (18)$$

$$\sigma_2 = \sqrt{\sigma_t^2 + \sigma^2} \quad (23)$$

as a result of convolution of  $G^{(1)}(\Delta t)$  and  $|G^{(1)}(\Delta t)|^2$  with the photomultiplier time response function, whose width is  $\sigma$ .

#### D. Results

Figure 1 shows the results of four experimental runs with amplified laser to chaotic intensity ratios of 0, .5, 1 and 2. No adjustments of any kind have been made to the data. The solid line in each case is a plot of Eq. (17) using the parameters of Table I which, except for  $R_0$ , were all determined independently of the experimental runs. In all cases the agreement between predicted and measured values is within 1%, indicating that the time correlation to be expected from amplified laser light is the same as that predicted for unamplified laser light of the same intensity as the amplified light, mixed with chaotic light whose intensity is that of the unavoidable spontaneous emission from the amplifier.

Joint Services Technical Advisory Committee  
F44620-C-74-0056

G. D. Blake and  
D. B. Scarl

#### REFERENCES

1. D. B. Scarl and S. R. Smith, "Time Correlations in Stimulated Emission," *Phys. Rev. A*10, pp. 709-713 (August 1974).
2. N. B. Abraham and S. R. Smith, "Stimulated Versus Spontaneous Emission as a Cause of Photon Correlations," *Phys. Rev. A*15, pp. 421-428 (January 1977).
3. E. B. Rockower, N. B. Abraham and S. R. Smith, "Evolution of the Quantum Statistics of Light," *Phys. Rev. A*17, pp. 1100-1112 (1 March 1978).
4. R. H. Dicke, "The Coherence Brightened Laser," in *Proc. of the 3rd International Conference on Quantum Electronics*, P. Grivet and N. Bloembergen (eds.), Columbia University Press, New York, 1964.
5. A. Kastler, "Le Caractere de Bosons des Photons et les Fluctuations d'un Faisceau Lumineux," in *Proc. of the 3rd International Conference on Quantum Electronics*, P. Grivet and N. Bloembergen (eds.), Columbia University Press, New York, 1964.
6. L. Mandel, "Stimulated Emission and Photon Correlations," *Phys. Rev. A*14, pp. 2351-2354 (December 1976).
7. D. B. Scarl and S. R. Smith, "Two-Photon Correlations in Light From a Gain Tube," *Phys. Rev. A*14, pp. 2235-2356 (December 1976).
8. J. H. van Vleck and D. L. Huber, "Absorption, Emission, and Linebreadths: A Semihistorical Perspective," *Rev. Mod. Phys.* 49, pp. 939-959 (October 1977).
9. J. U. White, "Very Long Optical Paths in Air," *J. Opt. Soc. Am.* 66, 411-416 (May 1976); *J. Opt. Soc. Am.* 32, 285 (1942).
10. R. J. Glauber, "Coherent and Incoherent States of the Radiation Field," *Phys. Rev.* 131, 2766-2788 (September 1963).

11. G. Lachs and D. R. Voltmer, "Photocount Time Interval Distribution for Superposed Coherent and Chaotic Radiation," *Journal of Applied Physics* 47, 346-349 (January 1976).
12. G. D. Blake, "Two-Photon Correlations in Amplified Laser Light," Dissertation submitted in partial fulfillment of the requirements for the Ph. D. degree (physics), Polytechnic Institute of New York, 1979.
13. D. B. Scarl, "Measurements of Photon Correlations in Partially Coherent Light," *Phys. Rev.* 175, 1661-1668 (November 1968).
14. G. Present and D. B. Scarl, "Two-Photon Correlations in a Mixture of Gaussian and Laser Light," *Applied Optics* 11, 120-124 (January 1972).
15. C. D. Cantrell and J. R. Fields, "Effect of Spatial Coherence on the Photoelectric Counting Statistics of Gaussian Light," *Phys. Rev. A* 7, 2063-2069 (June 1973).

TABLE I. Measured experimental parameters.

Ratio	$R_0$ cps	$I_0$ cps	$I_L$ cps	$I_N$ cps	$I_w$ cps	$\xi$
0:1	35,435	60321	0	1036	912	0
$\frac{1}{2}$ :1	32,523	62246	28491	1052	944	1.33
1:1	35,400	61277	60152	1309	930	1.28
2:1	50,485	60116	120233	1150	906	1.29

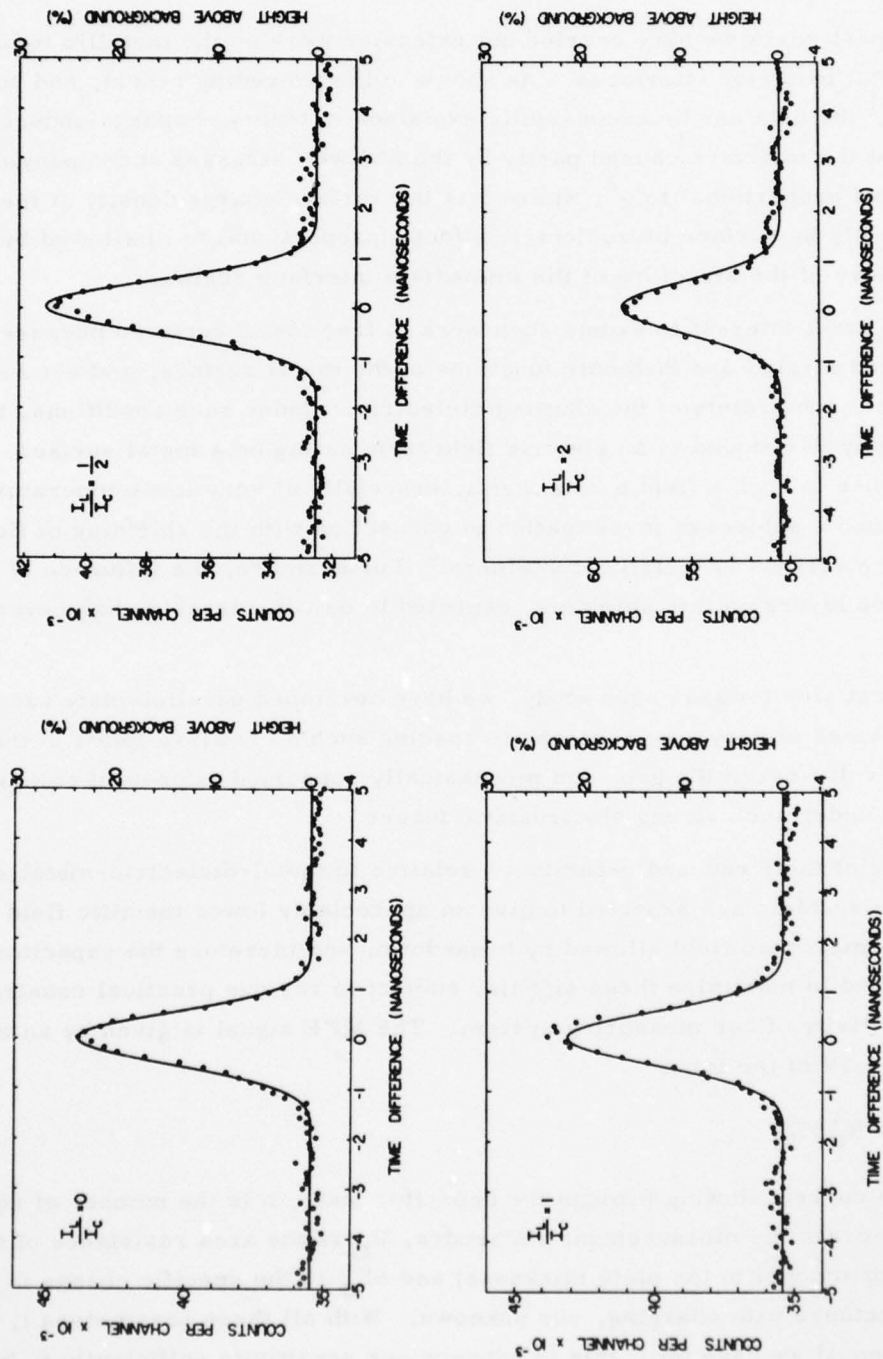


Fig. 1. Experimental counting rates and predicted curves for four runs in which the ratios of amplified laser light,  $I_L$ , to amplifier tube chaotic light in the same mode,  $I_C$ , were 0, .5, 1 and 2. The experimental statistical errors and the uncertainty in the predicted curve are each 1%.

## METALLIC FIELD EFFECT AT FREE METAL SURFACES

H.J. Juretschke, E. Segredo and M. Eschwei

In the past years we have carried out extensive work on the metallic field effect (MFE) at metal insulator interfaces. As shown in the preceding report, and some earlier ones,<sup>1</sup> the data can be successfully explained in terms of charge-induced stresses and strains at the interface caused partly by the Maxwell stresses accompanying electric fields (and proportional to  $q^2$ , where  $q$  is the surface charge density at the interface), and partly by surface piezoelectric effects (proportional to  $q$ ) allowed by the lower symmetry of the structure of the immediate interface region.

It is of great interest to extend such work to free metal surfaces because the charge-induced strains are then only functions of the metal surface, and not subject to any mechanical constraints of the abutting dielectric. Under such conditions, the system under study is reduced to an electric field terminating on a metal surface. The elastic response to such a field by the metal, especially at very low temperatures, has recently become a subject of investigation in connection with the shielding of fields of freely falling electrons in metallic containers.<sup>2</sup> Furthermore, the influence of the adsorbed surface layers on this shielding, expected to be considerable, has never been studied.

As a first step towards such study, we have developed parallel-plate vacuum capacitor structures of very small interplate spacing such as to allow fields of the order of  $10^5$  to  $10^6$  volts/cm in the gap, and mechanically supported to prevent collapse of the capacitor under such strong electrostatic forces.

Because of their reduced capacitance relative to metal-dielectric-metal sandwiches, these systems are expected to give an appreciably lower metallic field effect signal, at the maximum field allowed by breakdown, and therefore the capacitor plates were redesigned to maximize these signals, subject to various practical constraints on size and sensitivity of our measuring system. The MFE signal is given by an incremental voltage  $\delta V$  of the form

$$\delta V = -\ln R_{\square}^2 \delta \Sigma_{\square}$$

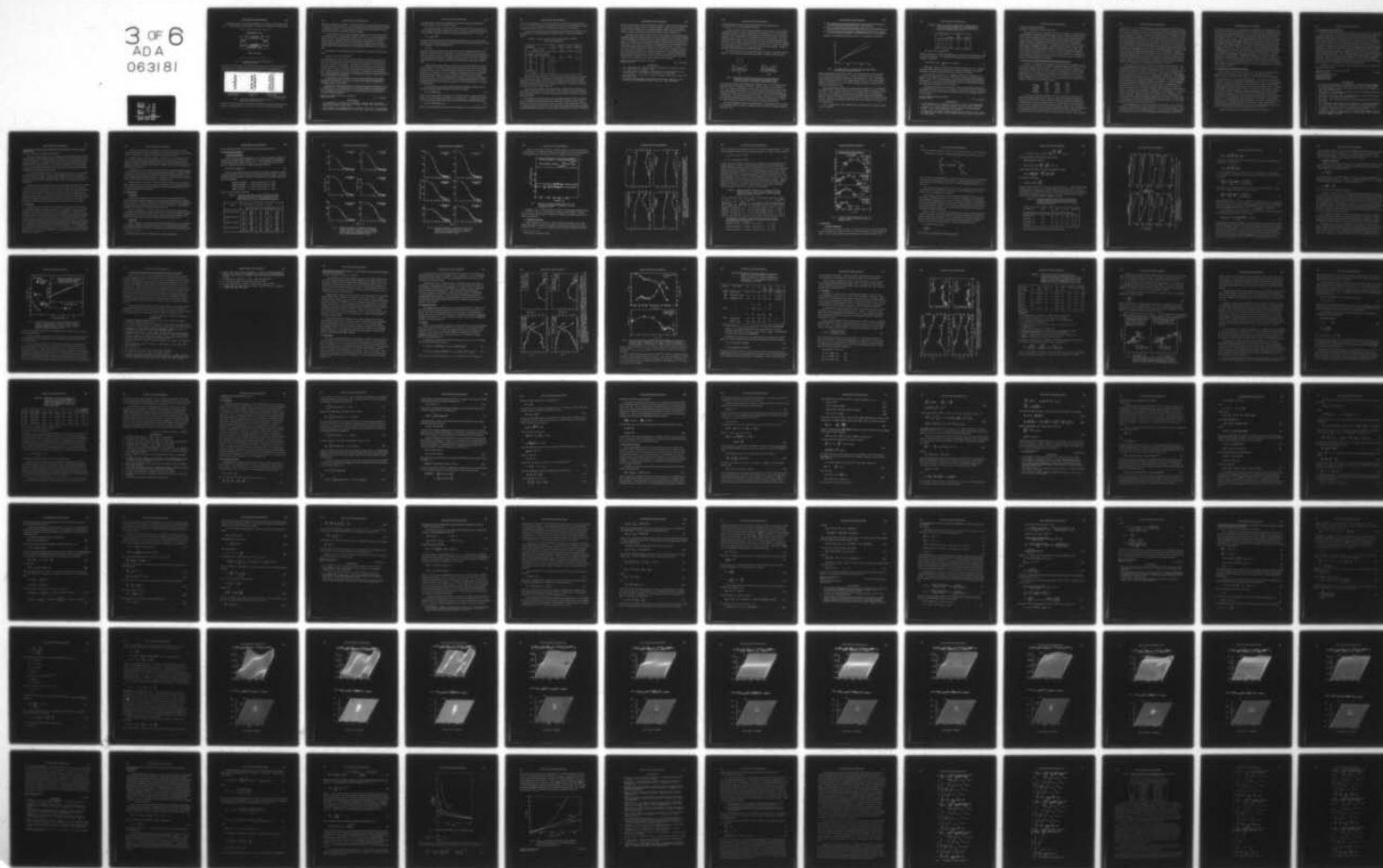
where  $I$  is the current flowing through the capacitor plate,  $n$  is the number of squares of the plate between the measurement electrodes,  $R_{\square}$  is the area resistance of the plate (inversely proportional to the plate thickness) and  $\delta \Sigma_{\square}$  is the specific change in the surface conductance with charging, our unknown. With all three parameters  $I$ ,  $n$ , and  $R_{\square}$  at our disposal we have been able to enhance our sensitivity sufficiently to be able to see the MFE under free surface conditions. We report here our first results on the effect of charge on silver electrodes exposed to the atmosphere.

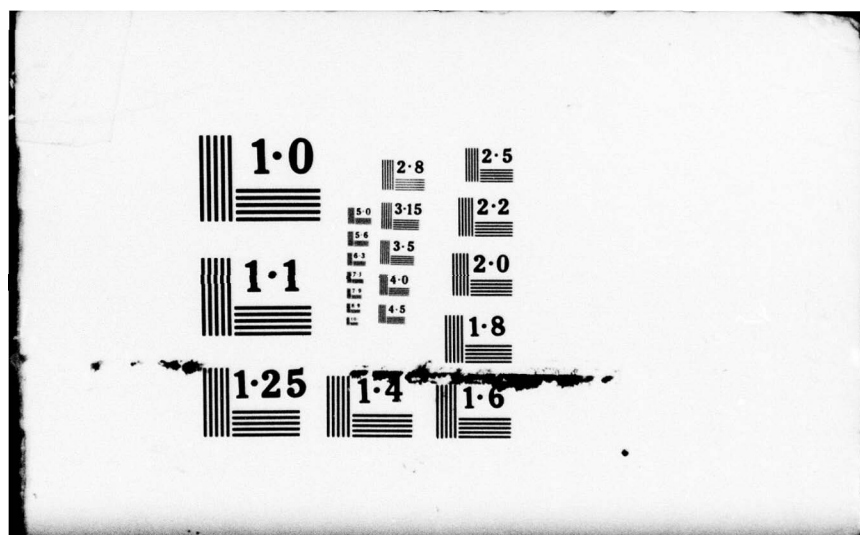
AD-A063 181

POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 9/3  
PROGRESS REPORT NUMBER 43 TO THE JOINT SERVICES TECHNICAL ADVIS--ETC(U)  
NOV 78 A A OLINER F44620-78-C-0074  
POLY-MRI-452.43-78 NL

UNCLASSIFIED

3 OF 6  
AD A  
063181





The depth profile of our sample arrangement is illustrated schematically in Figure 1. A silver sample, grown epitaxially on thin mica, is supported on optically flat

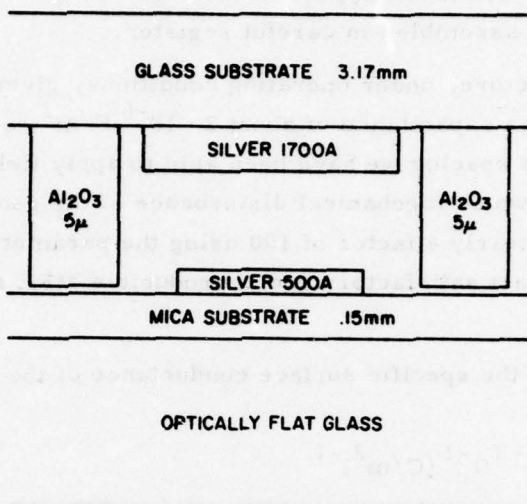


Fig. 1. Cross sectional view of vacuum-spaced field effect capacitor.

glass, and a counterelectrode of silver, on another glass plate, is kept at the desired separation by a set of spacers of alumina sputtered onto the second glass plate. In top view, the structure has the form shown in Figure 2. The thin electrode ribbon is

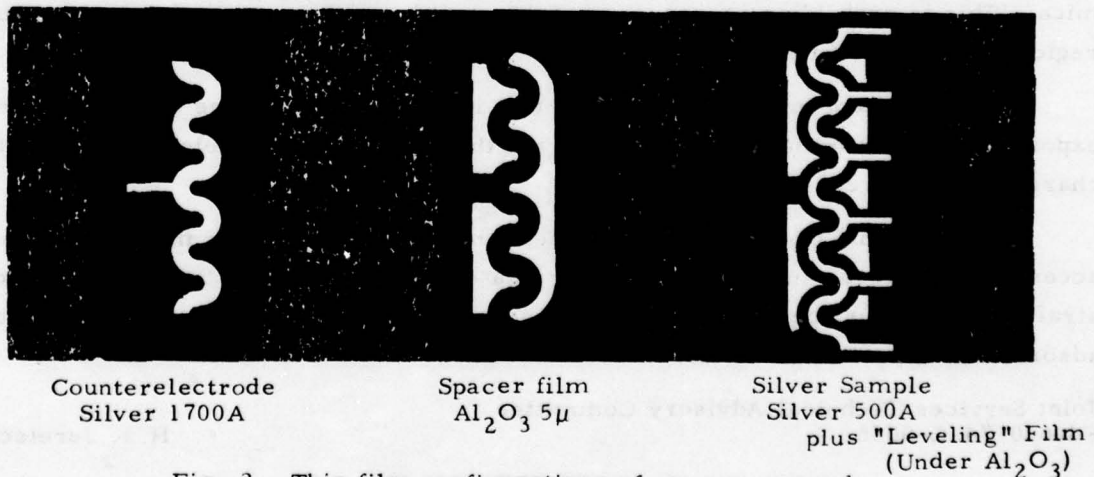


Fig. 2. Thin film configurations of vacuum-spaced field effect capacitor package.

in the form of connected half circles in order to be able to surround each portion of the sample by supporting spacers, as well as to increase the total length of the sample

between the measuring electrodes. Elaborate masks have been necessary for the  $\text{Al}_2\text{O}_3$  to permit all contact electrodes to leave the sample without introducing additional points of mechanical contact or electrical bypass between the top and bottom plates, and the whole structure must be assembled in careful register.

The complete structure, under operating conditions, gives a parallel plate vacuum capacitor with an area capacitance of about  $2 \cdot 10^{-6} \text{ F/m}^2$ , or a plate separation of about  $4 \cdot 10^{-6} \text{ m}$ . At this spacing we have been able to apply fields of more than  $10^7$  volts/m without breakdown or mechanical disturbance of the geometry, and with a sensitivity enhanced by nearly a factor of 100 using the parameters described above, we have been able to detect satisfactory and reproducible MFE signals on silver samples.

Our data give, for the specific surface conductance of the Ag/air interface, the value

$$\delta \Sigma_{\square} / \delta q = -3.0 \cdot 10^{-3} \Omega^{-1} (\text{C/m}^2)^{-1}$$

at room temperature. This number is smaller than that for Ag/mica by a factor of about five, indicating that the strain induced by a given surface charge density at the free silver surface is much smaller. The number is comparable to what we have observed for Ag/mylar, and Ag/kapton, suggesting that the bonding of Ag to these surfaces is weak. In terms of a simple model of surface piezoelectricity, this implies that a free silver surface is much less distortable by charging than when adhering to mica. This is probably a consequence of the equilibrium structure of the silver surface region in the different environments.

It must be kept in mind, of course, that in this experiment the silver surface, exposed to the atmosphere, is not clean, and that we are really seeing the effect of charging on a gas covered surface.

However, this preliminary experiment shows that the MFE is measurable on such accessible surfaces, and it opens the way to a systematic study of the charge-induced strain on metal surfaces that are clean or covered intentionally by various layers of adsorbed or deposited material.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

H.J. Juretschke

#### REFERENCES

1. H.J. Juretschke, D. Lischner and P. Mazumdar, "On the Origin of the Metallic Field Effect," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, p. 173 (1976).
2. J.M. Lockhart, F.C. Witteborn and W.M. Fairbank, "Evidence for a Temperature-dependent Surface Shielding Effect in Cu," Phys. Rev. Lett., Vol. 38, p. 1220 (1977).

## CHARGE-INDUCED SURFACE STRESSES AT METAL-INSULATOR INTERFACES

H.J. Juretschke, R. Boucarut and E. Segredo

The change in surface conductance of metals with electrostatic charging, the so-called metallic field effect, has been observed to contain two contributions, proportional to  $q$ , and  $q^2$ , respectively, where  $q$  is the surface charge density.<sup>1</sup> If we write

$$\delta \Sigma = -B_s q - B q^2 \quad (1)$$

then the coefficient  $B$  has found a full quantitative explanation as measuring the elastoresistance of the sample caused by strains induced in the sample by the Maxwell stresses in the dielectric. We can therefore write

$$B = \Sigma s \gamma \quad (2)$$

where the conductance  $\Sigma$  appears because the strains in the sample permeate it completely and make the  $q^2$ -term a bulk effect. The constant  $s$  measures the coupling between the strains and  $q$ , and  $\gamma$  is the effective elastoresistivity coefficient of the material for the given strain configuration. Experimental data on the various known constants and the measurement of  $B$  give satisfactory agreement between both sides of Equation (2).<sup>2</sup>

Largely because of the similarity, as well as characteristic differences, observed in the behavior of  $B$  and  $B_s$  in ferromagnetic metals under the variation of the magnetic parameters, we have concluded that the coefficient  $B_s$  must have a similar interpretation as  $B$ . However, since it is independent of the thickness of the metal sample,  $B_s$  must be related strictly to a change of surface conductance caused by surface stresses linear in  $q$ . This change presumably originates in charge-induced changes of the surface scattering of current carriers,<sup>3</sup> but at least phenomenologically it can also be described in terms of the change of conductance of a thin surface layer as a function of surface strain.

If a surface stress  $X_s$  causes strains throughout an effective thickness  $t_s$ , then we can formulate this model in analogy with Eq. (2), and obtain

$$B_s q = (\sigma t_s)(S X_s / t_s) \gamma_s \quad (3)$$

where  $\sigma$  is the conductivity of the metal,  $S$  is the metal's elastic constant for isotropic stress, and  $\gamma_s$  is the surface counterpart of  $\gamma$  of Equation (2). Combining Eqs. (2) and (3), we can use the measured values of  $B$  and  $B_s$  to determine the surface stress  $X_s$ :

$$X_s / q = t(s/S)(B_s/B)(\gamma / \gamma_s) \quad (4)$$

where  $t$  is the sample thickness, introduced to eliminate  $\Sigma = \sigma t$ . If we further assume

that the strain dependence of scattering is the same in bulk and in the surface layer, i.e.,  $\gamma_s = \gamma$ , then all quantities on the right side of Eq. (4) can be obtained from experiment, and this gives a measure of the stress per unit charge in the surface.

The result of combining the measurements of  $B_s$  and  $B$  with the constants  $s$  and  $S$  characteristic of various substrates and metals yields the values of  $X_s/q$  shown in Table I.

TABLE I. Surface stresses  $X_s/q$  of metal-dielectric interfaces.  
(Units of  $(N/m)/(C/m^2)$ )

Substrate:		Mica	Mylar	Glass	Kapton
$s[\text{in } (C/m^2)^{-2}]$		.054	.86	2.6	.77
Metal	$S(10^{-11} \text{ m/N})$				
Ag	1.28	4.3	-0.35	-0.40	-2.3
Au	1.28	-1.3	2.3		6.1
Ag/Au	1.3	-0.074			
Cu	0.87	18			
Al	1.02	1.3			
Sb	1.01	7.9			
NiFe 90-10	0.54	2.0			
80-20	0.54	1.5			

The central result of this table is that all surface stresses obtained in this manner are close to the order of magnitude  $(-1(N/m)(C/m^2)^{-1})$  predicted by theoretical estimates.<sup>4</sup> In more detail, however, there are significant and surprising variations of this stress between different metal/substrate combinations, ranging over more than one power of ten, and showing either sign.

At this time there exists no theory with which to compare these results for different metals and substrates, but they appear to offer sufficient variety to indicate that the method proposed here can be used to study interface stresses caused by charging in quite some detail.

For example, the result for the Ag-Au alloy falls between the values for pure Ag and Au, and indicates that there exists one composition at which the charge-induced surface stresses at the mica interface vanish, at least at room temperature. In terms of a model of surface piezoelectricity, this implies that this particular interface retains its centrosymmetric character so that piezoelectric effects have to be absent. Similarly,

the large stress value for the Cu/mica interface implies a particularly strong response to electrostatic charging, which is of interest whenever copper is used as a shield. Finally, the great variation observed between different metal/substrate combinations appears to make it impossible to extrapolate the data to other combinations.

It should be kept in mind, however, that while the surface layer thickness  $t_s$  does not appear explicitly in the final result, Eq. (4), its magnitude is nevertheless pertinent to the reasonableness of the above analysis. Measurements in the Au-Ag/mica system, using a very thin layer of Ag, indicate the thickness of penetration of the strains to be of the order of  $25 \text{ \AA}$ . Hence the model of a two-parallel-layers system applies strictly only to those metals whose carriers have a mean free path of order  $t_s$  or less, i.e., the two Ni-Fe alloy compositions of Table I. In all other cases, conduction in the surface region has to be treated on a more microscopic basis, but this should primarily affect the equality of the ratio  $\gamma_s/\gamma$ . Since the range of  $\gamma$ 's encountered among widely differing metals is relatively small, such a deviation should not be large enough to upset the relative magnitudes of the deduced stresses shown in Table I.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

H. J. Juretschke

#### REFERENCES

1. H. J. Juretschke and L. Goldstein, "Nonlinear Response in the Metallic Field Effect," Phys. Rev. Lett., 29, 767 (1972).
2. D. Lischner and H. J. Juretschke, "Maxwell Stresses at Charged Surfaces from Elastoresistance," to be submitted for publication in J. Appl. Phys.
3. A. Berman and H. J. Juretschke, "Origin of the Metallic Field Effect," Phys. Rev., B11, 2903 (1975).
4. C. Herring, "Gravitationally Induced Electric Field Near a Conductor, and Its Relation to the Surface-Stress Concept," Phys. Rev., 171, 1361 (1968).

## ELASTORESISTANCE AND ELECTRON TUNNELING IN OXIDIZED IRON

T. Pignataro, H.J. Juretschke and M. Eschwei

As part of a study of the microwave properties of metal surfaces covered with inhomogeneous oxides, we have attempted to build prototype structures of two metallic regions connected by an oxide of the kind likely to be found under natural oxide growth conditions, and to measure their room-temperature electrical properties. We chose iron as the metal, because it and its oxides are an often-encountered practical metallic surface, even though its conventional MOM junctions, obtained by thin film overlays, are largely ohmic at room temperature.<sup>1</sup> This is attributed to the significant conductivity of the pertinent oxides, or to a very low barrier at the oxide interface. We have therefore concentrated on different MOM geometries.

The most successful structure has been obtained using the configuration shown in Figure 1. An iron film of about  $1000 \text{ \AA}$  is deposited on a substrate, and a thin line

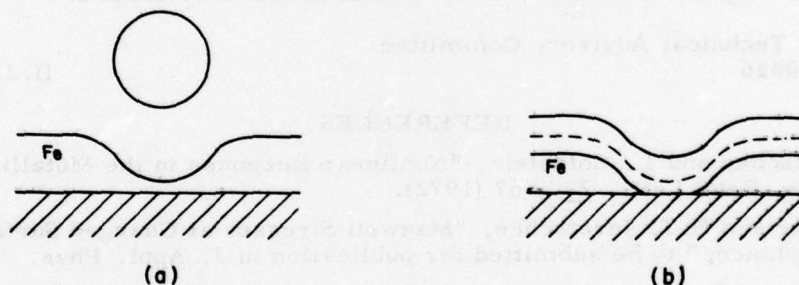


Fig. 1. Method of preparing thin-line junctions of  $\text{FeO}_x$  between two iron films. (a) Thin-line deposit in the shadow of a wire mask; (b) Schematic oxide growth in the thin-line region.

across the film is produced by shadowing the deposition with a  $10^{-3} \text{ cm}$  diameter wire stretched about  $10^{-2} \text{ cm}$  above the substrate. This line, shown in cross section in Fig. 1(a), is optically transparent, but electrically continuous, and of very low resistance. If the sample is then exposed to a wet atmosphere at elevated temperature, an oxide grows all along film surface, and forms a layer of the form shown schematically in Figure 1(b). This geometry produces an  $\text{Fe-FeO}_x\text{-Fe}$  junction of a cross sectional area significantly smaller than obtained in the usual film overlay structure. It therefore tends to magnify the influence of the junction in transport along the film.

The room temperature electrical measurements on samples prepared in this manner show three important changes in properties relative to those of the initial configuration of Fig. 1(a):

- (1) The resistance increases significantly faster with oxidation than expected by the uniform thinning of the sample because of the oxide skin.
- (2) The I-V characteristic becomes nonlinear, with  $V(-I) = -V(I)$ , at current values that become progressively smaller as the sample resistance increases.
- (3) The strain sensitivity of resistance  $(1/R)(dR/de)$  increases with increasing  $R$ , and also shows a nonlinearity, becoming smaller as the current increases.

A typical I-V characteristic is shown in Fig. 2, for a sample containing three

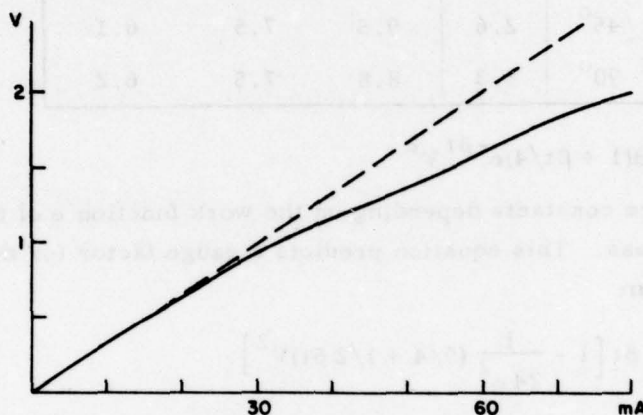


Fig. 2. I-V characteristic of a lightly oxidized sample having three thin-line junctions in series.

lightly oxidized thin-line junctions in series. Its initial shape has the form expected for current tunneling through a barrier. At higher currents, heating of the sample causes deviations from the ideal exponential tunneling behavior. Also, some samples show a slight opening of the I-V loop, suggesting that some type of switching between different current states, such as is found in amorphous materials, may play a role in conduction through the barrier.

The elastoresistance response of a sample similar to that of Fig. 2 is summarized in Table I below. Clearly, the presence of the junctions enhances the effect of strain, in all cases. The pronounced decrease in the strain gauge factor  $(1/R)(dR/de)$  with large currents is in qualitative agreement with the behavior expected from tunneling junctions. However, the change in angular dependence of the gauge factor with increased currents, indicating an appreciable anisotropy in the elastic behavior of the junction, is not yet understood.

The I-V characteristic of the simplest model of a square barrier junction is given at low currents by<sup>2</sup>

TABLE I. Strain gauge factor  $(1/R)(dR/de)$  for oxidized thin iron films: (a) of uniform thickness, (b) containing three thin-line junctions in series.  $\theta$  is the angle of the applied strain relative to the current.

$\theta$	(a)	(b)		
		5 mA	10 mA	15 mA
$0^\circ$	4.0	11.3	8.3	6.0
$45^\circ$	2.6	9.5	7.5	6.1
$90^\circ$	1.3	8.8	7.5	6.2

$$I/V = Ae^{-\beta t} + B(1 + \beta t/4)e^{-\beta t}V^2 \quad (1)$$

where A, B and  $\beta$  are constants depending on the work function  $\phi$  of the barrier, and  $t$  is the barrier thickness. This equation predicts a gauge factor for the junction, at constant I, of the form

$$(1/V)(dV/de) = \beta t \left[ 1 - \frac{1}{24\phi^2} (9/4 + 1/2 \beta t) V^2 \right] \quad (2)$$

This result contains the observed decrease with increasing I (or V). It predicts a primarily linear growth of the gauge factor with barrier thickness, and it allows estimating an absolute magnitude of the strain sensitivity of a junction.

For a  $\text{Fe-FeO}_x\text{-Fe}$  junction,  $\phi$  is expected to be below 1 eV, and then  $\beta$  is a fraction of  $1 \text{ \AA}$ . Hence a barrier with  $t \sim 5 \text{ \AA}$  would be compatible with the results of Table I. In fact, it is very likely that  $\beta$  is much smaller and  $t$  much larger than this first estimate. At this point the most important conclusion is that the barrier model gives magnitudes and trends of the correct order.

The prediction that the strain sensitivity is enhanced appreciably by thicker barriers suggests that such junctions on the surfaces of oxidized iron and other metals may contribute to the relatively large strain dependence of the surface impedance observed on such surfaces.<sup>3</sup>

Air Force Office of Scientific Research  
AFOSR-76-3030B

H.J. Juretschke

#### REFERENCES

1. J.E. Christopher, R.V. Coleman, A. Isin and R.C. Morris, "Experiments with Tunnel Junctions Using Ferromagnetic Metals," *Phys. Rev.* **172**, p. 485 (1968).
2. R. Stratton, "Volt-Current Characteristics for Tunneling Through Insulating Films," *J. Phys. Chem. Solids*, **23**, p. 1177 (1962).
3. H. Bertoni and L.M. Silber, "Strain Dependence of Rust Layers at X-Band," Progress Report No. 42 to JSTAC, Polytech, Inst. of New York, Report No. R-452.42-77, p. 170 (1977).

## NEW SOLID STATE MATERIALS

E. Banks, S. Nakajima, R. Sacks and M. Shone

Research in the area of new materials is devoted to synthesis, crystal growth and characterization of materials having interesting electrical, optical and magnetic properties which may provide a basis for new optical, electro-optic and magneto-optic devices. Section A describes the structural and luminescence properties of the new complex fluorides of divalent Sm and Eu with Mg, their behavior in dilution with  $\text{SrMgF}_4$ , and the characterization of the striking energy transfer from  $\text{Eu}^{2+}$  to  $\text{Sm}^{2+}$  in solid solutions of  $\text{SmMgF}_4$  in  $\text{EuMgF}_4$ . Section B gives the result of a single crystal X-ray analysis of a tetragonal mixed fluoride of composition  $\text{K}_{0.54}\text{Mn}_{0.5}\text{Fe}_{0.54}\text{F}_3$ . Section C discusses the properties of a new series of solid solutions of  $\text{LiInP}_2\text{O}_7$  in cubic  $\text{ZrP}_2\text{O}_7$  in which up to 20 mole of  $\text{Li}^+$  ions have been substituted, making the material a promising candidate for a solid electrolyte in batteries based on Li metal, for applications requiring high energy density.

A. Fluorescence in New Divalent Rare Earth Magnesium Fluorides

Last year, we reported on the synthesis and some luminescent properties of newly discovered ternary fluorides of composition  $\text{EuMgF}_4$ ,  $\text{SmMgF}_4$  and  $\text{SrMgF}_4$ , and some of the solid solutions formed among them. The unit cells of all three compounds are orthorhombic, similar to the unit cells of  $\text{BaMnF}_4^1$  and  $\text{BaMgF}_4^2$ , the latter two being piezoelectric at room temperature and showing non-linear optical behavior such as frequency doubling. An apparent anomaly in the unit cell parameters as compared to those of  $\text{BaMgF}_4$  has been resolved by a more careful analysis of the X-ray data on our materials. The unit cells are compared below:

	<u>a(A)</u>	<u>b(A)</u>	<u>c(A)</u>
$\text{BaMgF}_4$	5.81	14.509	4.125
$\text{SmMgF}_4$	5.661	14.440	3.965
$\text{EuMgF}_4$	5.658	14.430	3.933
$\text{SrMgF}_4$	5.637	14.459	3.917

In contrast to the barium compounds, which melt congruently, and thus can be obtained as sizeable single crystals, all three of these new compounds decompose on melting (near  $900^\circ\text{C}$ ) into the binary fluorides. We had previously reported difficulty in preparing the Sm and Eu compounds directly from the difluorides, necessitating the use of Mg vapor on mixtures of the trifluorides and  $\text{MgF}_2$ . This has now been found to be a result of the presence of some oxide in the starting materials, which was eliminated by preheating the starting materials ( $\text{EuF}_3$  and  $\text{SmF}_3$ ) with ammonium fluoride. Attempts to grow single crystals have been unsuccessful thus far because of the incongruent

melting behavior. Flux growth experiments, using molten  $\text{MgCl}_2$  and  $\text{SrCl}_2$ , are now being set up for  $\text{SrMgF}_4$  as a model compound and host crystal for  $\text{EuMgF}_4$  and  $\text{SmMgF}_4$ . In the absence of single crystals, we have been attempting to elucidate the crystal structure from powder X-ray data. Assuming the published structure<sup>1</sup> of  $\text{BaMnF}_4$ , structure factor calculations on  $\text{EuMgF}_4$  were carried out and compared with those derived from the observed powder X-ray intensities. While satisfactory agreement was not obtained, the results suggested that the structures are related. A test for second harmonic generation was negative, indicating that  $\text{EuMgF}_4$  may be centrosymmetric and suggesting the possibility that the structure may be that of the hypothetical high-temperature form of  $\text{BaMgF}_4$  and that  $\text{EuMgF}_4$  may undergo a low-temperature transition to a ferroelectric phase. We are now setting up the measurement of the  $^{151}\text{Eu}$  Mössbauer spectrum which may yield information about the existence of such a transition. A large sample of  $\text{SrMgF}_4$  is being prepared for a study of the structure by neutron diffraction, using the new method<sup>3</sup> of line profile analysis for structure refinement on polycrystalline samples. Attempts to prepare thin films of  $\text{EuMgF}_4$  by vacuum evaporation failed because of the decomposition of the samples in the evaporation vessel.

The luminescence properties of these new compounds are quite unusual. The compound  $\text{EuMgF}_4$  displays a bright blue fluorescence under 365 nm excitation. The blue emission occurs in a broad blue band peaking at 437 nm. The excitation spectrum is also a broad band with a maximum at 350 nm. In solid solution with  $\text{SrMgF}_4$ , the maximum efficiency is found at 75 mole %  $\text{EuMgF}_4$ , but the efficiency of pure  $\text{EuMgF}_4$  is only about 5% (relative) less than that of the brightest solid solution. This indicates the near-absence of concentration quenching in this material. In contrast, the brightness of solid solutions of  $\text{SmMgF}_4$  in  $\text{SrMgF}_4$  is a maximum at 10 mole %  $\text{SmMgF}_4$ , indicating substantial concentration quenching. This behavior of  $\text{EuMgF}_4$  places the material in the class of "stoichiometric" phosphors where the active luminescent species can be present in very high concentration without reducing the luminescent efficiency. Examples of such phosphors are  $\text{NdP}_5\text{O}_{14}$ <sup>4-6</sup> and  $\text{LiNdP}_4\text{O}_{12}$ .<sup>7,8</sup> The latter materials have been made into miniaturized solid state lasers, emitting in the near infrared (1.06  $\mu\text{m}$ ). The highly efficient broadband blue emission of  $\text{EuMgF}_4$  and its solid solutions with  $\text{SrMgF}_4$  and  $\text{BaMgF}_4$  suggests that lasers tunable in the blue region of the spectrum can be made from single crystals of these materials, by controlling the optical feedback under flashlamp excitation.

Pure  $\text{SmMgF}_4$  fluoresces a weak red under 365 nm excitation, but its excitation maximum is near 470 nm; the emission spectrum is a series of sharp lines in the red and near IR, with the strongest peak near 680 nm. In solid solution in  $\text{SrMgF}_4$ , maximum efficiency is found at 10 mole %  $\text{SmMgF}_4$ , indicating substantial concentration

quenching, as stated above. In solid solutions of composition  $\text{Eu}_{1-x}\text{Sm}_x\text{MgF}_4$ , the fluorescence under 365 nm excitation shows no  $\text{Eu}^{2+}$  emission for Sm concentrations above 10 mole %, and very little Eu emission at 1 mole % Sm, even though excitation is near the maximum for  $\text{Eu}^{2+}$ . This indicates very efficient energy transfer from excited  $\text{Eu}^{2+}$  centers to  $\text{Sm}^{2+}$  centers. The complete absence of the  $\text{Eu}^{2+}$  spectrum in samples containing more than 1% Sm suggests that the major mechanism of energy transfer is non-radiative (resonance transfer). Measurements of the lifetime of the excited state of the  $\text{Eu}^{2+}$  have confirmed that this is occurring. The excited state lifetime in pure  $\text{EuMgF}_4$  is about  $10^{-7}$  sec. This is reduced to  $10^{-9}$  sec (for the  $\text{Eu}^{2+}$  emission) when 1%  $\text{SmMgF}_4$  is in solid solution. This substantial shortening of the fluorescent lifetime is strong evidence for non-radiative energy transfer from  $\text{Eu}^{2+}$  to  $\text{Sm}^{2+}$ . An experiment in which a mixed powder of  $\text{EuMgF}_4$  and  $\text{SmMgF}_4$  was irradiated with 365 nm radiation also showed excitation of the  $\text{Sm}^{2+}$  red emission, but the  $\text{Eu}^{2+}$  emission was also present. This indicates that radiative energy transfer is possible, but does not play a major role in the solid solution materials, which do not show the  $\text{Eu}^{2+}$  emission. This indicates that the non-radiative transfer rate is very rapid compared to the excited state lifetime of the  $\text{Eu}^{2+}$ .

#### B. Mixed Alkali Transition Metal Fluorides

Work in this area has been deferred in favor of the research discussed above. A crystal structure analysis has been completed on one member of the series  $\text{K}_x\text{Mn}_x\text{Fe}_{1-x}\text{F}_3$ , nominally  $\text{K}_{0.5}\text{Mn}_{0.5}\text{Fe}_{0.5}\text{F}_3$ , whose composition was determined crystallographically to be  $\text{K}_{0.54}\text{Mn}_{0.5}\text{Fe}_{0.04}^{2+}\text{Fe}_{0.5}^{3+}\text{F}_3$ . The results are summarized as follows: A specimen of flux-grown  $\text{K}_{0.54}(\text{Mn}, \text{Fe})\text{F}_3$  was studied at room temperature by single crystal X-ray diffractometry. The crystal has a structure of the tetragonal tungsten bronze type with a doubled  $c$  axis. The unit cell parameters are  $a = 12.765(1) \text{ \AA}$ ,  $c = 8.002(1) \text{ \AA}$ , and the space group is  $\text{P}_2^{4bc}$ . Least-squares refinement was carried out with 2395 symmetry-independent reflections collected with an automatic diffractometer (Mo- $\text{K}\alpha$  radiation). The final R value ( $\sum |F_o^2 - kF_c^2| / \sum |F_o^2|$ ) is 0.049, with anisotropic thermal parameters. In the structure, potassium atoms fully occupy pentagonal (CN=15) sites and partially occupy tetragonal (CN=12) sites. The transition metal ions occupy three different kinds of octahedra. The mean M-F distances in each octahedron are:  $2.100 \text{ \AA}$  for M(1)-F,  $1.939 \text{ \AA}$  for M(2)-F, and  $1.995 \text{ \AA}$  for M(3)-F. These distances indicate that the M(1) sites are mainly occupied by bivalent ions, chiefly  $\text{Mn}^{2+}$ , the M(2) sites by  $\text{Fe}^{3+}$  and the M(3) sites by  $\text{Mn}^{2+}$  and  $\text{Fe}^{3+}$ .

A paper<sup>9</sup> on this work has been submitted to Acta Crystallographica.

### C. Potential New Solid Electrolytes

We have continued in the attempt to introduce large concentrations of alkali metals into cubic  $\text{ZrP}_2\text{O}_7$ , whose structure contains large interconnected cavities making it a candidate for application as a solid electrolyte in batteries based on alkali metals. Previously we had reported the introduction of up to 10 mole %  $\text{Li}^+$  ions, compensated by  $\text{Y}^{3+}$  replacing  $\text{Zr}^{4+}$  ions. Recently we have succeeded in introducing up to 20 mole % Li into this compound. This was accomplished by using indium as the charge-compensating ion. Compositions formulated as  $\text{Zr}_{1-x}\text{Li}_x\text{In}_x\text{P}_2\text{O}_7$  were prepared. Samples with values of  $x$  from 0 to 0.2 show only the pattern of cubic  $\text{ZrP}_2\text{O}_7$ , with line shifts indicating solid solution formation. At higher concentrations ( $x$  values) a new phase appears, apparently  $\text{LiInP}_2\text{O}_7$ . This composition appears to be a previously unreported compound. The powder diffraction pattern appears to be orthorhombic from a preliminary study. We are now preparing larger samples for study by nuclear magnetic resonance to determine whether the lithium ions are mobile, and to determine whether ceramic samples can be prepared for conductivity measurements.

Continuation of work in this area has been funded for two years by a grant from the Army Research Office.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

E. Banks

U.S. Army Research Office  
DAAG-29-75-C-0096  
DAAG-29-78-G-0105  
DAAG-29-78-G-0150

### REFERENCES

1. E.T. Keve, S.C. Abrahams and J.L. Bernstein, "Crystal Structure of Pyroelectric Barium Manganese Fluoride,  $\text{BaMnF}_4$ ," J. Chem. Phys., 51, 4928 (1969).
2. J.G. Bergman and G.R. Crane, "Non-Linear Optical Properties of  $\text{BaMgF}_4$ ," J. Appl. Phys., 46, 4645 (1975).
3. H.M. Rietvelt, "A Profile Refinement Method for Nuclear and Magnetic Structures," J. Appl. Cryst., 2, 65-71 (1969).
4. H.P. Weber, T.C. Damen, G.H. Danielmeyer and B.C. Tofield, "Nd-Ultraphosphate Laser," Appl. Phys. Lett., 22, 534 (1973).
5. H.P. Weber, "Nd Pentaphosphate Lasers," Optics and Quantum Electronics, 7, 431 (1975).
6. H.G. Danielmeyer, "Stoichiometer Laser Materials," in Festkörperprobleme XV, Adv. in Solid State Physics (New York: Pergamon Press, 1975), pp. 253-277.
7. T. Yamada, K. Otsuka and J. Nakano, "Fluorescence in Lithium Neodymium Tetrphosphate Single Crystals," J. Appl. Phys., 45, 5096 (1974).
8. K. Otsuka, T. Yamada, M. Saruwatari and T. Kimura, "Spectroscopy and Laser Oscillation Properties of Lithium Neodymium Tetrphosphate," IEEE J. Quantum Electronics, QE-11, 330 (1975).

## ULTRASONIC AND MICROWAVE DIELECTRIC RELAXATION OF LIQUID DIALKYL CARBONATES

D. Saar, J. Brauner, H. Farber and S. Petrucci

Ultrasonic absorption data for pure dimethyl, diethyl, dipropyl, dibutyl and propylene carbonates in the frequency range of 3-300 MHz and in the temperature range of 25-55°C are reported. An ultrasonic relaxation process of the Debye type with a relaxation frequency  $f_R = 8.5$ -10 MHz at 25°C, which is largely independent of the length of the alkyl chain in the alkoxy group, has been found for the non-cyclic carbonates. No such process has been observed for the cyclic propylene carbonate and for dimethoxymethane which lacks the carbonyl group.

The observed relaxation process is interpreted as being due to a cis-trans isomerization of the alkoxy groups. By using Lamb's approach for thermal relaxation processes both the activation parameters  $\Delta H_r^\ddagger$  and  $\Delta S_r^\ddagger$  for the reverse step of the equilibrium and the thermodynamic parameters  $\Delta H^\circ$  and  $K$  for the equilibrium have been determined.

Complex dielectric permittivities in the microwave frequency range of 0.6-67 GHz for the above non-cyclic carbonates at 25°C are also reported. The data can be described over almost all the frequency range studied by a single Debye relaxation process. The high end of the frequency range shows some positive deviations of the loss in accord with literature data. The dielectric relaxation process is interpreted as due to a segmental motion of the molecule under the influence of the electric field. The different nature of the relaxation mechanisms for the ultrasonic and dielectric processes is attributed to the different type of perturbing function inherent in the two methods.

A. Introduction

Mechanical waves of ultrasonic radio frequencies have been used in the past to investigate the dynamics of molecular relaxation in pure liquids.<sup>1</sup> Similarly, electromagnetic waves at microwave frequencies have been employed to study the decay of polarization in these same substances due to the relaxation of the orientation of dipolar molecular groups.<sup>2</sup> Very rarely have both types of waves been employed by the same research group to investigate the molecular dynamics of organic liquids, the most common situation being mutual ignorance of the information obtainable by the alternate tool. Because of the different natures of the perturbing waves complementary information of great value can be obtained by the use of both techniques. The compressions and rarefactions caused by the mechanical waves can disturb a pressure sensitive equilibrium which has a  $\Delta V_s (= \Delta V_T - \theta \Delta H / \rho C_p) \neq 0$  as indicated by the well-known relation  $(\partial \ln K / \partial p)_s = -\Delta V_s / RT$ .<sup>3,4</sup>

Thus, the concentration of a molecular form which is small relative to some other form at low pressure may be substantially increased at high pressure if its molar volume is smaller or its enthalpy is higher than that of the other form.

An applied electric field, for example, in the form of an electromagnetic wave, may also cause a shift in the equilibrium between two molecular forms. If the two forms have a different dipole moment the equilibrium constant relating them can be changed according to the relation  $(\partial \ln K / \partial E) = \Delta M / RT$ , where  $\Delta M$  is the change in the dipole moment per mole.<sup>4</sup>

Alternatively, if a given molecular form with a non-zero dipole moment (because of the presence of one or more polar groups) always predominates, a dielectric relaxation process associated with the orientations (following the alternating field) of one or more of these polar groups will be present. In this latter case the dielectric relaxation process\* may not be related to the ultrasonic relaxation process.

Dialkyl carbonates were chosen as the starting point because of their relation to esters which contain one of the two alkoxy groups present in the carbonates and because of the possibility that these molecules might be considered model monomers for the corresponding polycarbonate polymers.

## B. Experimental

### 1. Apparatus

The ultrasonic equipment and procedures have been described elsewhere.<sup>5</sup> A linear least squares method was applied to all the absorption (dB) data versus distance at each frequency. This improved the precision of the calculated absorption coefficients and gave correlation coefficients better than  $r^2 = 0.99$  in all instances. The dielectric equipment and procedures have been described elsewhere.<sup>6</sup>

Cannon viscometers No. 0 and No. 1 (Cannon, University Park, Pennsylvania) with manufacturer calibration certificates were used. Two 50 ml pycnometers calibrated with distilled water at 25.0°C were also used. Thermostating of the ultrasonic and dielectric cells and of the viscometers and pycnometers was with  $\pm 0.05^\circ\text{C}$ .

### 2. Materials

Diethyl-, dipropyl- and dibutylcarbonates (Eastman Kodak) were vacuum distilled in an all glass apparatus and used shortly thereafter. Dimethoxymethane (lab stock) was distilled at atmospheric pressure. Propylene carbonate (Eastman Kodak) was used

---

\* Associated with the dipolar orientation of the predominant species.

without further purification, the result being only of a qualitative nature.

### C. Results and Calculations

#### 1. Ultrasonic Relaxation

Figure 1 shows the results expressed as  $\alpha/f^2$  versus the frequency  $f$  (MHz) for dimethyl carbonate and diethyl carbonate at the various temperatures investigated.  $\alpha$  is the sound absorption coefficient ( $\text{neper cm}^{-1}$ ). The solid lines represent values given by a Debye type function for a single relaxation process:<sup>3,4</sup>

$$\alpha/f^2 = (A/1 + (f/f_R)^2) + B \quad (1)$$

Figure 2 shows similar plots for the other two carbonates investigated. Table I collects the quantities  $A$ ,  $B$ ,  $f_R$  and the sound velocities for these systems. The sound velocities are linear functions of temperature expressible by the equations calculated by linear regression:

$$\text{dimethyl carbonate: } u = 1196 - 3.67(t-25) \text{ m/s } r^2 = 0.996$$

$$\text{diethyl carbonate : } u = 1179 - 3.47(t-25) \text{ m/s } r^2 = 0.985$$

$$\text{dipropyl carbonate: } u = 1122 - 4.23(t-25) \text{ m/s } r^2 = 0.995$$

$$\text{dibutyl carbonate : } u = 1253 - 3.20(t-25) \text{ m/s } r^2 = 0.988$$

where  $r$  is the correlation coefficient.

TABLE I. Ultrasonic parameters  $A$ ,  $B$  ( $\text{cm}^{-1} \text{sec}^2$ ) and  $f_R$  (MHz) according to Eq. (1) for the liquid carbonates investigated at 25, 40 and 55°C. Corresponding sound velocities (m/sec) for the same systems and temperatures.

liquid	$t^\circ, \text{C}$	$A \times 10^{17}$ $\text{cm}^{-1} \text{sec}^2$	$B \times 10^{17}$ $\text{cm}^{-1} \text{sec}^2$	$f_R$ MHz	$u$ meter $\text{sec}^{-1}$
Dimethyl carbonate	25	4500	56	8.5	1198
	40	2930	70	15	1137
	55	2230	70	22	1088
Diethyl carbonate	25	2357	43	9.5	1176
	40	1570	30	17	1135
	55	1100	50	30	1072
Dipropyl carbonate	25	1600	50	10	1219
	40	1140	60	15	1163
	55	850	60	24	1092
Dibutyl carbonate	25	1540	60	8.5	1256
	40	1100	70	12	1199
	55	750	60	20	1160

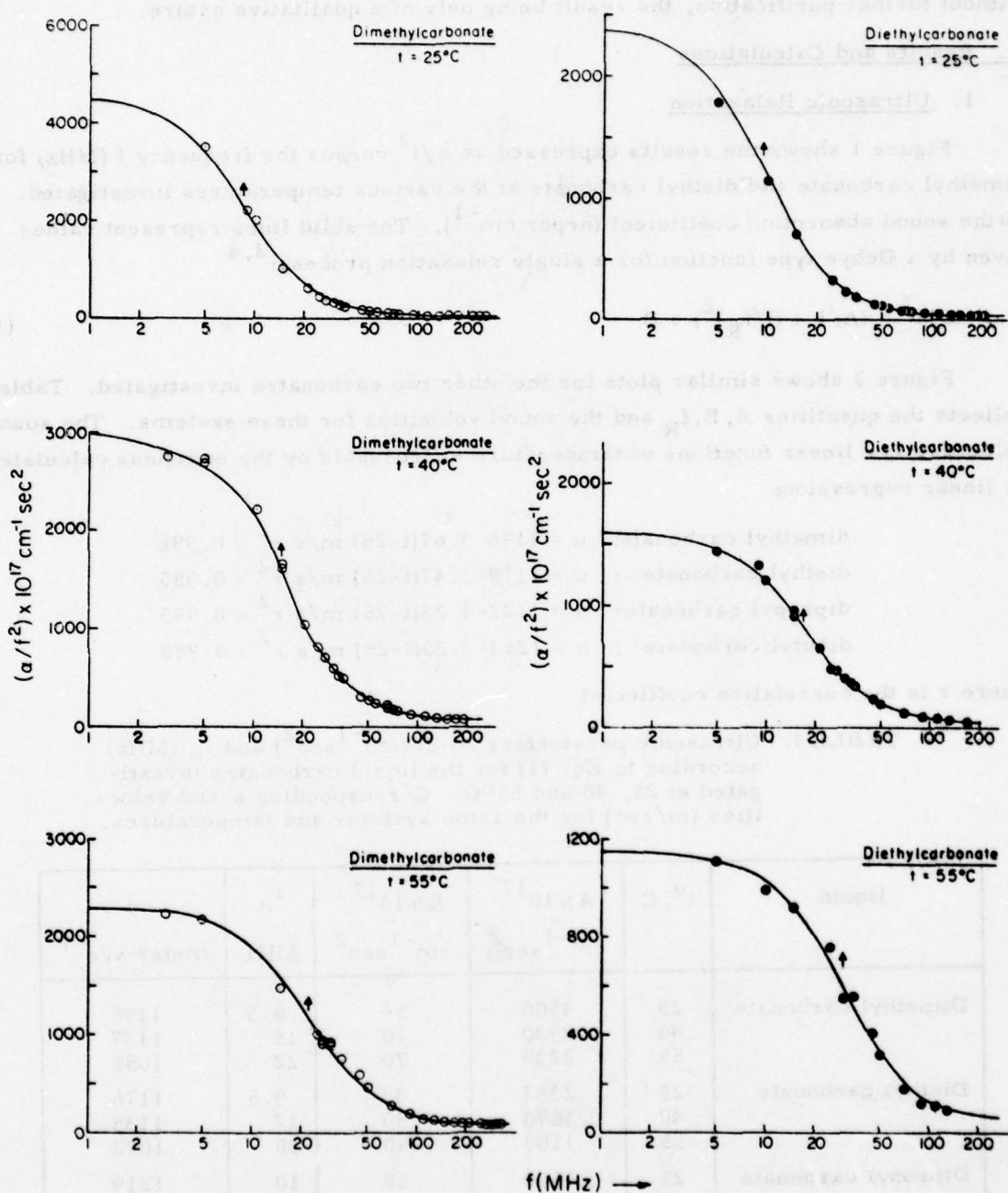


Fig. 1. Ultrasonic relaxation of dimethyl carbonate and diethyl carbonate at 25, 40 and 55°C. Ordinate:  $\alpha/f^2 10^{17} \text{ cm}^{-1} \text{ sec}^2$ . Abscissa: frequency  $f \text{ (MHz)}$ . The solid lines are the calculated values according to a single Debye relaxation function.

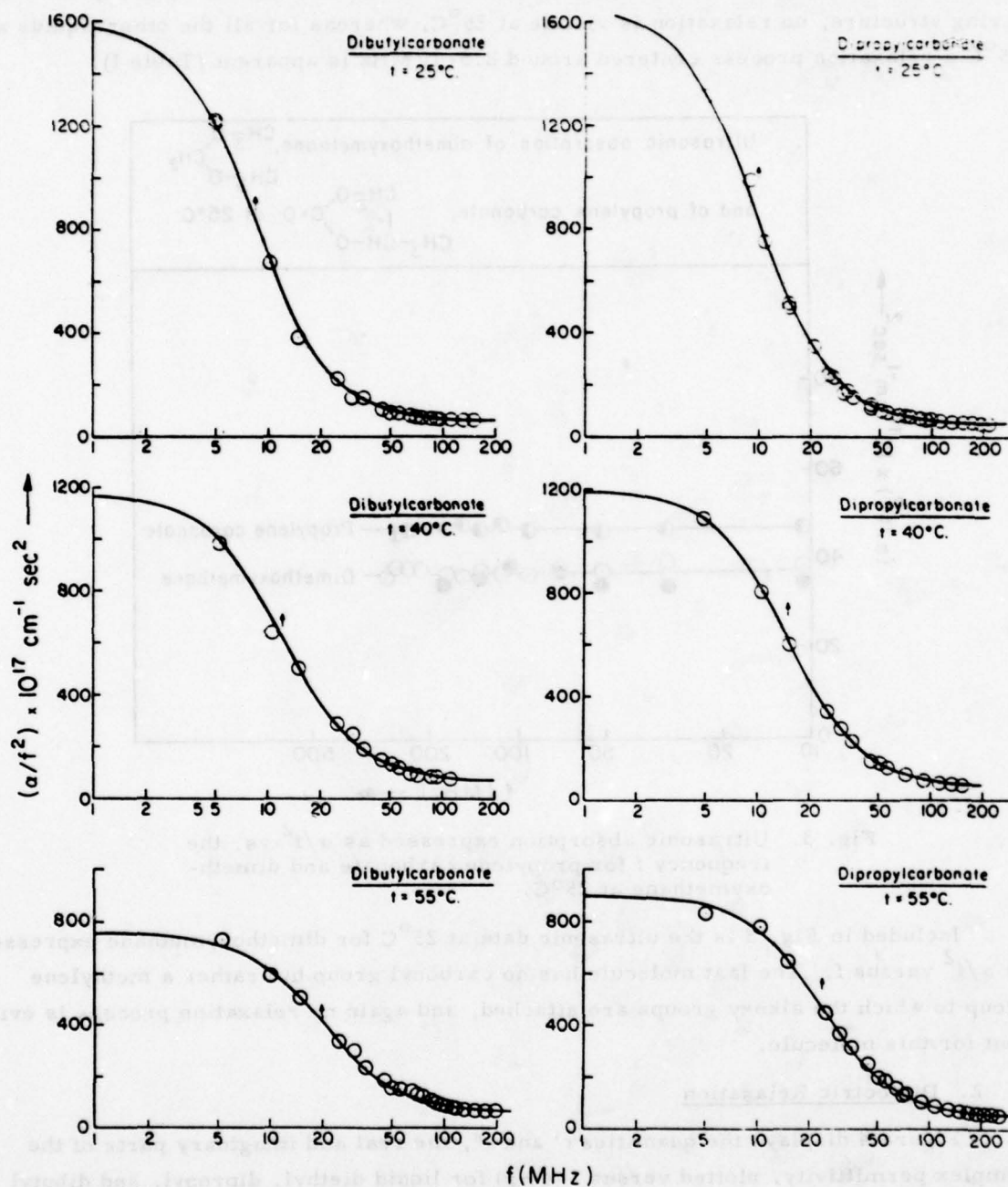


Fig. 2. Ultrasonic relaxation of dipropyl carbonate and dibutyl carbonate at 25, 40 and 55°C. The solid lines are the calculated values according to a single Debye relaxation function.

Figure 3 shows that for propylene carbonate which has its alkoxy groups held in a ring structure, no relaxation is visible at 25°C, whereas for all the other liquids at 25°C a relaxation process centered around 8.5-10 MHz is apparent (Table I).

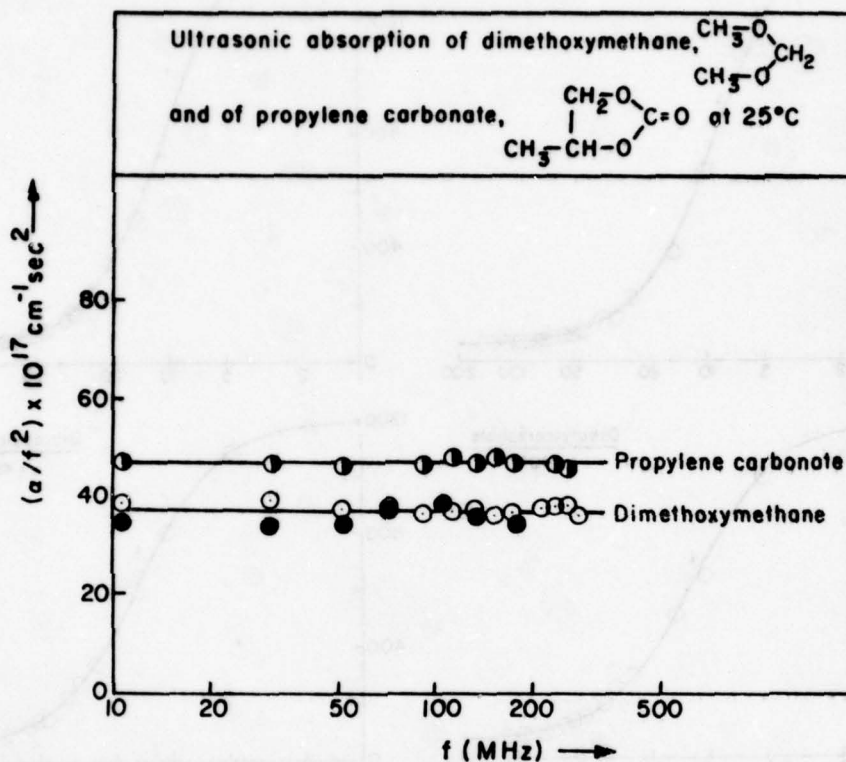


Fig. 3. Ultrasonic absorption expressed as  $\alpha/f^2$  vs. the frequency  $f$  for propylene carbonate and dimethoxymethane at 25°C.

Included in Fig. 3 is the ultrasonic data at 25°C for dimethoxymethane expressed as  $\alpha/f^2$  versus  $f$ . The last molecule has no carbonyl group but rather a methylene group to which the alkoxy groups are attached, and again no relaxation process is evident for this molecule.

## 2. Dielectric Relaxation

Figure 4 displays the quantities  $\epsilon'$  and  $\epsilon''$ , the real and imaginary parts of the complex permittivity, plotted versus  $f$  (GHz) for liquid diethyl, dipropyl, and dibutyl carbonates at 25°C. The data for dimethyl carbonate\* has been reported previously.<sup>6</sup>

\* Shown here for comparison sake.

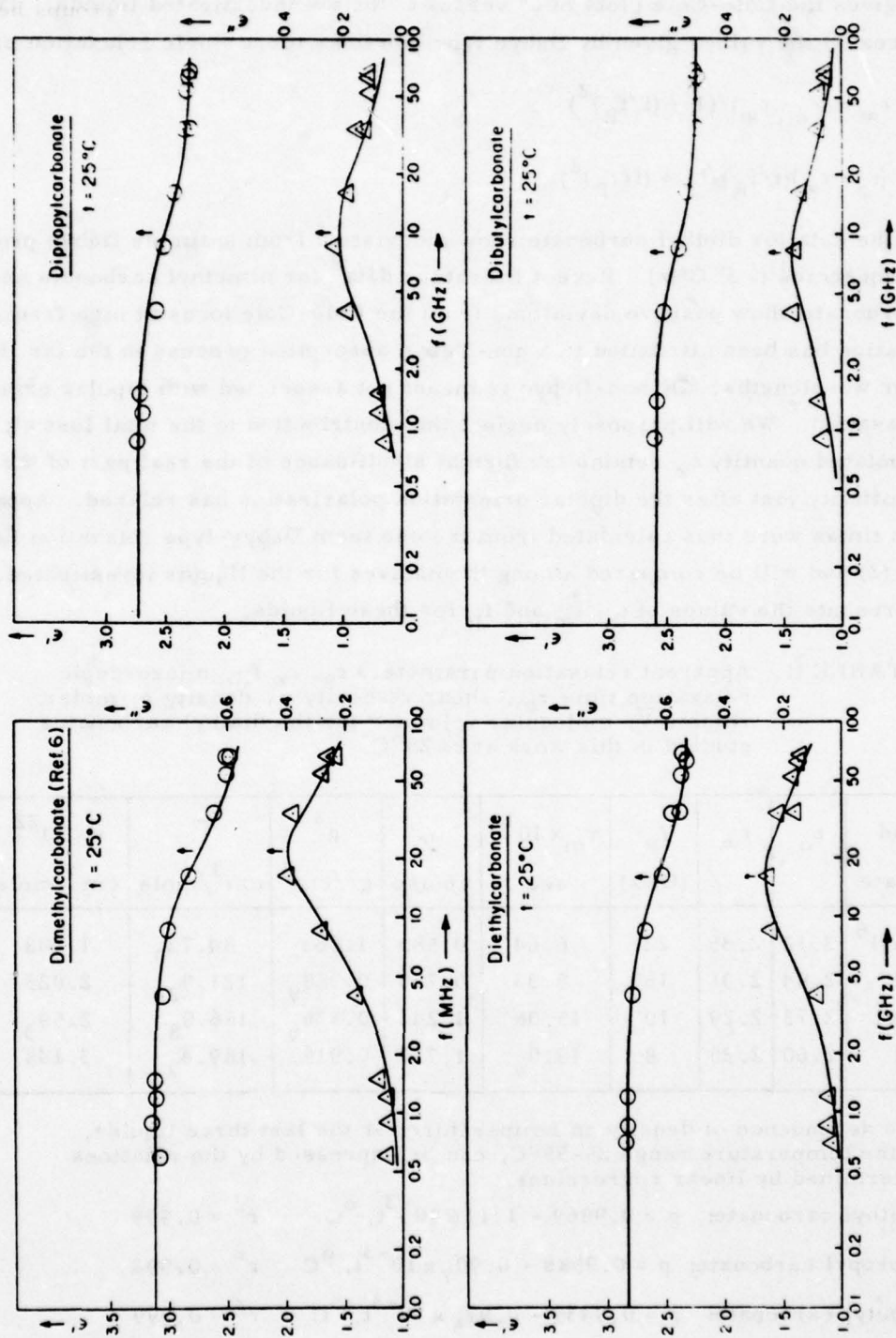


Fig. 4. Real part  $\epsilon'$  and coefficient of the imaginary part  $\epsilon''$  of the complex permittivity plotted vs. the frequency  $f$  (GHz) for dimethyl, diethyl, dipropyl and dibutyl carbonates at 25°C. The solid lines are calculated functions according to single Debye relaxation function.

Figure 5 gives the Cole-Cole plots of  $\epsilon''$  versus  $\epsilon'$  for the investigated liquids. The solid lines represent the values given by Debye type functions for a single relaxation process<sup>7</sup>

$$\begin{aligned}\epsilon' &= \epsilon_{\infty} + (\epsilon_0 - \epsilon_{\infty}) / (1 + (f/f_R)^2) \\ \epsilon'' &= (\epsilon_0 - \epsilon_{\infty})(f/f_R) / (1 + (f/f_R)^2)\end{aligned}\quad (2)$$

In Fig. 4 the data for diethyl carbonate show a deviation from a simple Debye process at high frequencies ( $> 35$  GHz). Recent literature data<sup>8</sup> for dimethyl carbonate and diethyl carbonate show positive deviations from the Cole-Cole locus at high frequencies. This deviation has been attributed to a non-Debye absorption process in the far IR at millimeter wavelengths. By non-Debye is meant not associated with dipolar orientational relaxation. We will purposely neglect this contribution to the total loss  $\epsilon''$ ; thus the extrapolated quantity  $\epsilon_{\infty}$  retains the formal significance of the real part of the complex permittivity just after the dipolar orientation polarization has relaxed. Apparent relaxation times were thus calculated from the one term Debye-type relaxation functions Eq. (2) and will be compared among themselves for the liquids investigated. Table II presents the values of  $\epsilon_0$ ,  $\epsilon_{\infty}$  and  $f_R$  for these liquids.

TABLE II. Apparent relaxation parameters  $\epsilon_0$ ,  $\epsilon_{\infty}$ ,  $f_R$ , microscopic relaxation time  $\tau_m$ , shear viscosity  $\eta$ , density  $\rho$ , molar volume  $\bar{V}$ , molecular volume  $v$  for the dialkyl carbonates studied in this work at  $t = 25^\circ\text{C}$ .

Liquid carbonate	$\epsilon_0$	$\epsilon_{\infty}$	$f_R$ (GHz)	$\tau_m \times 10^{12}$ sec	$\eta$ cpoise	$\rho^*$ gr/cm <sup>3</sup>	$\bar{V}$ cm <sup>3</sup> /mole	$v \times 10^{22}$ cm <sup>3</sup> /mole
Dimethyl <sup>6</sup>	3.12	2.35	22	6.64	0.585	1.063 <sub>0</sub>	84.73 <sub>9</sub>	1.408
Diethyl	2.84	2.31	16	9.33	0.750	0.968 <sub>9</sub>	121.9 <sub>2</sub>	2.025
Dipropyl	2.73	2.29	10	15.06	1.243	0.936 <sub>6</sub>	156.0 <sub>8</sub>	2.59 <sub>3</sub>
Dibutyl	2.60	2.25	8	19.0 <sub>0</sub>	1.717	0.919 <sub>5</sub>	189.4 <sub>9</sub>	3.148

\* The dependence of density on temperature for the last three liquids, in the temperature range  $25-55^\circ\text{C}$ , can be expressed by the relations determined by linear regressions.

Diethyl carbonate:  $\rho = 0.9969 - 1.11 \times 10^{-3}t, ^\circ\text{C}$   $r^2 = 0.999$

Dipropyl carbonate:  $\rho = 0.9586 - 0.90_7 \times 10^{-3}t, ^\circ\text{C}$   $r^2 = 0.993$

Dibutyl carbonate:  $\rho = 0.9439 - 0.97_5 \times 10^{-3}t, ^\circ\text{C}$   $r^2 = 0.999$

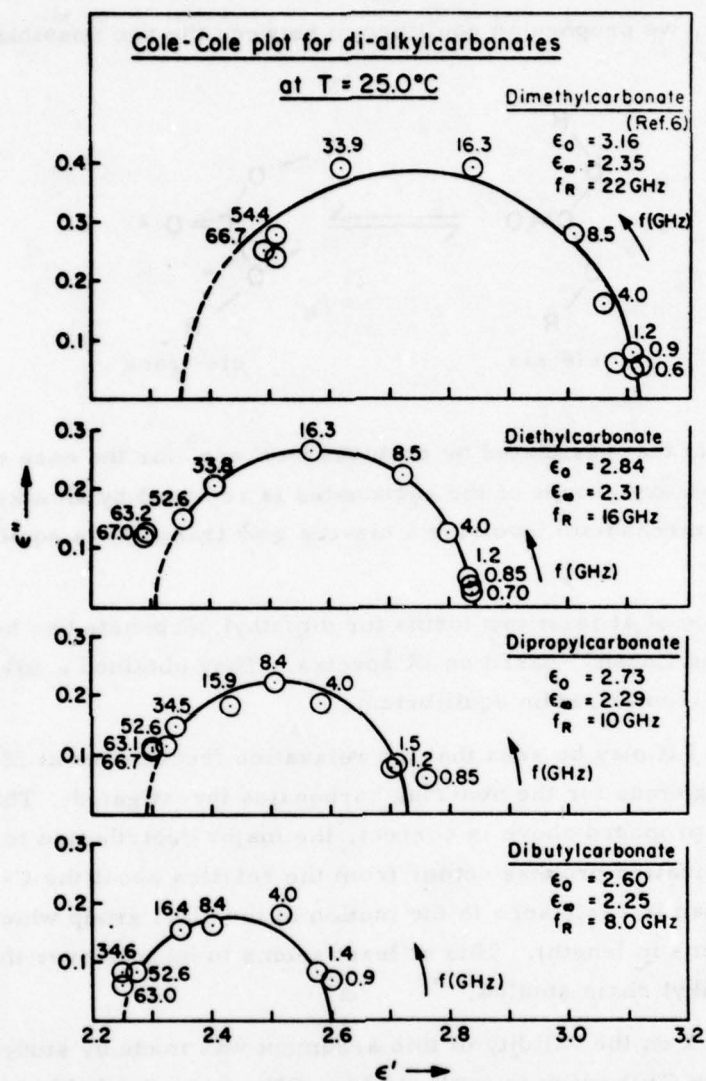


Fig. 5. Cole-Cole plots of the quantity  $\epsilon''$  vs.  $\epsilon'$  for dimethyl, diethyl, dipropyl and dibutyl carbonates at  $25^\circ\text{C}$ .

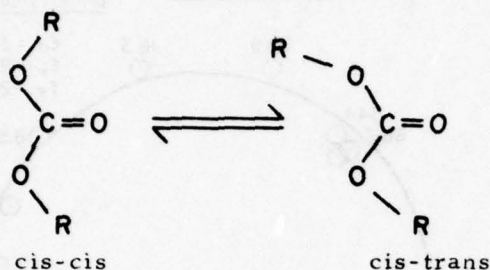
#### D. Discussion

##### 1. Ultrasonic Relaxation

Of the liquid carbonates investigated only propylene carbonate has its alkoxy groups largely immobilized by a ring structure; only for this carbonate is a relaxation process absent. Therefore the hypothesis is advanced that the molecular mechanism of the

ultrasonic relaxation process involves a cis-trans isomerization of the alkoxy groups.

Specifically, we propose an equilibrium between the two possible structures



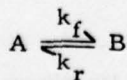
which is similar to that postulated by previous workers<sup>9</sup> for the case of the alkyl esters where one of the alkoxy groups of the carbonates is replaced by an alkyl group. (A concerted type of mechanism involving a cis-cis  $\rightleftharpoons$  trans-trans equilibrium is not ruled out.)

The existence of at least two forms for dimethyl carbonate has been reported recently by Katon and Cohen<sup>10</sup> based on IR spectra. They obtained a  $\Delta H = 2.6 \pm 0.5$  kcal/mole for the isomerization equilibrium.

From Table I it may be seen that the relaxation frequencies at 25°C are of the same order of magnitude for the non-ring carbonates investigated. This means that, if the equilibrium proposed above is correct, the major contribution to the energy barrier of the isomerization process comes from the rotation about the C-O bond (and not from the hydrodynamic resistance to the motion of the alkyl group which goes from one to four carbon atoms in length). This at least seems to be true over the variation in the length of the alkyl chain studied.

Another check on the validity of this argument was made by studying dimethoxymethane. Here the C=O group is replaced by a CH<sub>2</sub> group (probably causing the C-O bonds to become true single bonds). This should cause more molecular flexibility and permit free rotation of the alkoxy groups; indeed at 25°C no ultrasonic relaxation process is visible over the entire frequency range investigated (Figure 3).

Given the molecular process



with  $K = k_f/k_r$  one can write the following equations:<sup>9</sup>

$$\tau^{-1} = 2\pi f_R = k_f + k_r = k_r(1 + K) = \frac{kT}{h} e^{\frac{\Delta S_r^\ddagger}{R}} e^{\frac{-\Delta H_r^\ddagger}{RT}} (1 + K) , \quad (3)$$

where the symbols used have their usual meanings.

Rearranging<sup>11</sup> we recover:

$$\ln(\tau^{-1}/T) = \left[ \ln \frac{k}{h} + \frac{\Delta H_r^\ddagger}{R} \right] - \frac{\Delta H_r^\ddagger}{RT} + \ln(1 + K) .$$

A plot of  $\ln(\tau^{-1}/T)$  versus  $1/T$  yields a straight line of slope:

$$\text{slope} = \frac{d \ln(\tau^{-1}/T)}{d(1/T)} = - \frac{\Delta H_r^\ddagger}{R} - \frac{K}{K+1} \frac{\Delta H^0}{R} \quad (4)$$

and showing as intercept:

$$\text{intercept} = \ln \left[ \frac{k}{h} + \frac{\Delta S_r^\ddagger}{R} \right] .$$

Plots of the quantity  $\ln(\tau^{-1}/T)$  versus  $1/T$  are shown in Figure 6. The calculated values of  $\Delta S_r^\ddagger$  are given in Table III. To calculate  $\Delta H_r^\ddagger$  one needs to know  $\Delta H^0$  as is obvious from Equation (4). This latter parameter is calculable from the maximum excess sound absorption per wavelength,  $\mu_{\max} = (A/\lambda)u f_R$ . The values appearing on the right hand side of the above equation have been tabulated in Table I.

TABLE III. Activation parameters  $\Delta S_r^\ddagger$ ,  $\Delta H_r^\ddagger$  and thermodynamic parameters  $\Delta H^0$  and equilibrium constant  $K$  for the isomeric equilibria in liquid dialkyl carbonate studied by ultrasonic relaxation.

Liquid carbonate	$\Delta S_r^\ddagger$ e. u. (= cal/mole deg.)	$\Delta H_r^\ddagger$ Kcal/mole	$\Delta H^0$ Kcal/mole	$K \times 10^3$
Dimethyl-	-4.6	5.5	2.8 <sub>5</sub>	8.1
Diethyl-	-0.1 <sub>1</sub>	6.8	3.7 <sub>2</sub>	0.1 <sub>9</sub>
Dipropyl-	-6.0	5.0	3.0 <sub>2</sub>	6.1
Dibutyl-	-6.8	4.8	2.0 <sub>1</sub>	33.8

Still following Lamb,<sup>9</sup> for the process  $A \rightleftharpoons B$  one has:

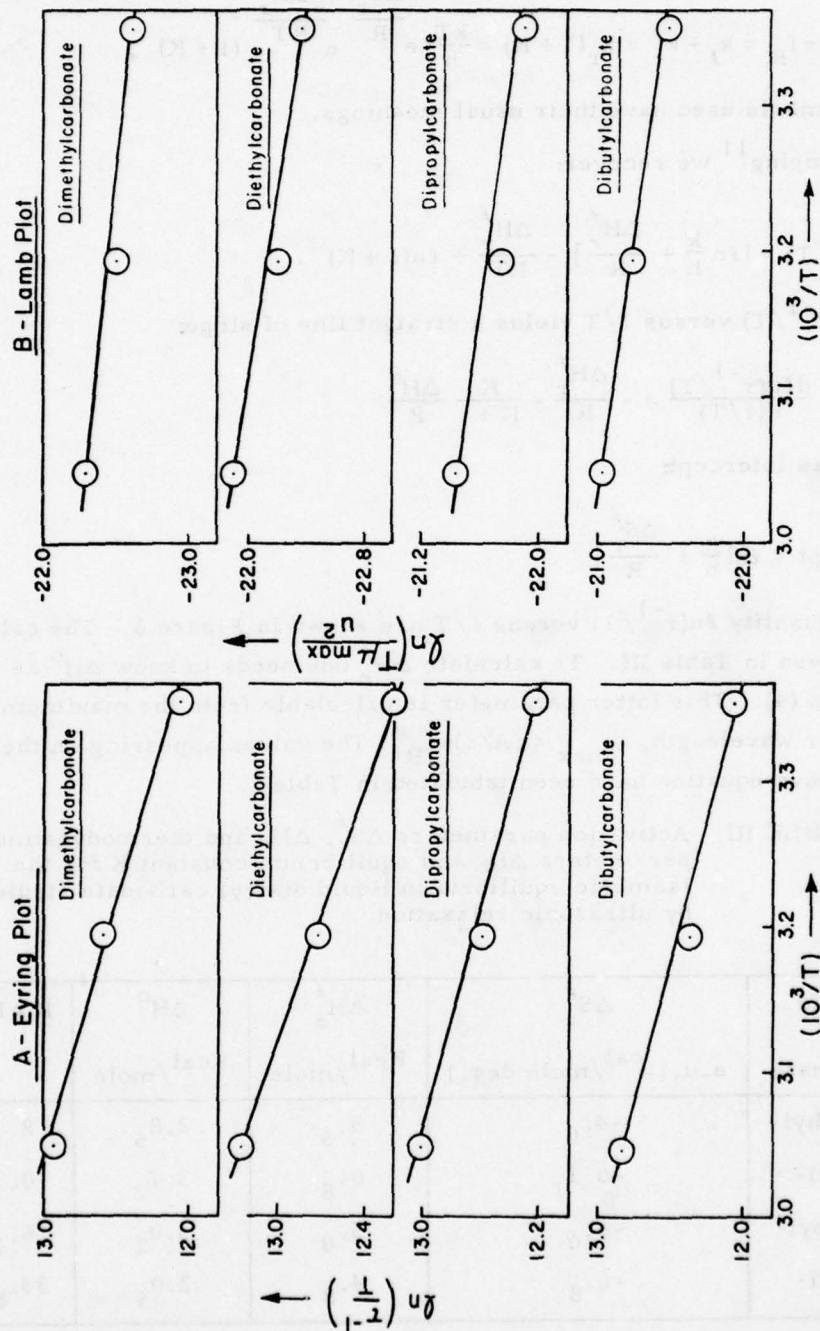


Fig. 6. (A) Eyring plot of  $\ln(\tau^{-1}/T)$  vs.  $1/T$  for the liquid carbonates investigated;  $\tau$  is the ultrasonic relaxation time.  
 (B) Lamb plot of  $\ln(\mu_{\max} T/u^2)$  vs.  $1/T$  for the liquid carbonates investigated;  $\mu_{\max}$   $\tau A/2$  ufr (see Table I).

$$\mu_{\max} = \frac{\pi}{2\beta_s} \frac{\Delta V_s^2}{RT} \left( \frac{1}{[A]} + \frac{1}{[B]} \right)^{-1},$$

introducing  $K = [B]/[A]$ , and if the concentrations of A and B are expressed in mole fraction units,<sup>9</sup>  $[A] + [B] = 1$ :

$$\mu_{\max} = \frac{\pi}{2\beta_s} \frac{\Delta V_s^2}{RT} \frac{K}{(1+K)^2} \quad (5)$$

It is now assumed that  $\Delta V_T \approx 0$  (see Ref. 9) for an isomerization reaction so that  $\Delta V_s \approx -\theta \Delta H^0 / \rho c_p$ . Substituting this approximation for  $\Delta V_s$  and  $\beta_s = 1/\rho u^2$  into Eq. (5) one obtains:

$$\frac{T\mu_{\max}}{\rho u^2} = \frac{\pi}{2} \frac{\theta^2}{\rho^2 c_p^2} \frac{(\Delta H^0)^2}{R} \frac{e^{-\Delta H^0/RT}}{(1 + e^{-\Delta H^0/RT})^2} \quad (6)$$

where again for an isomerization reaction it has been assumed<sup>9</sup> that  $\Delta S^0 \approx 0$  so that  $\Delta G^0 \approx \Delta H^0$ . Modifying slightly Eq. (6) one arrives at:

$$\frac{\mu_{\max}}{Tu^2} = \frac{\pi}{2} \frac{R\theta^2}{\rho^2 c_p^2} \left( \frac{\Delta H^0}{RT} \right)^2 \frac{e^{-\Delta H^0/RT}}{(1 + e^{-\Delta H^0/RT})^2} \quad (7)$$

The function  $x^2 e^{-x} / (1 + e^{-x})^2$ , with  $x = \Delta H^0 / RT$ , has a maximum at  $x = 2.4$ . This means that for  $x > 2.4$  or  $\Delta H^0 > 2.4 RT$ , the quantity  $\mu_{\max} / Tu^2$  will increase by decreasing  $x$  or, in other words, by increasing  $T$ . This behavior has been verified for all the carbonates investigated which implies that  $\Delta H^0 > 1.42 \text{ kcal/mole}$  at  $T = 298.15^\circ \text{K}$ .

Now, in Eq. (6) one may approximate<sup>9</sup>  $(1 + e^{-\Delta H^0/RT})^2 \approx 1$  and, taking the natural log of both sides of the equation, one may write:

$$\ln(T\mu_{\max}/u^2) = \ln\left[\left(\frac{\pi}{2} \frac{\theta^2}{\rho^2 c_p^2}\right) \left(\frac{\Delta H^0}{R}\right)\right] - \Delta H^0/RT \quad (8)$$

Neglecting<sup>9</sup> the temperature dependence of the first term on the right, one would expect a plot of  $\ln(T\mu_{\max}/u^2)$  versus  $1/T$  to be linear with a slope  $= -\Delta H^0/R$ . This is actually the case as seen from Figure 6. The values of  $\Delta H^0$  obtained are presented in Table III. For dimethyl carbonate  $\Delta H^0 = 2.85 \text{ kcal/mole}$  which agrees within the experimental error with the value of  $2.6 \pm 0.5 \text{ kcal/mole}$  obtained by infrared spectra.<sup>10</sup> The agreement is fairly remarkable considering the approximations involved in Equation (8). Table III also reports the values of  $K = e^{-\Delta H^0/RT}$ . It may be seen that  $K \ll 1$  in all

cases, showing the presence of one of the two molecular configurations in large excess with respect to the other. Furthermore, the values of  $\Delta H_r^\ddagger$  are comparable for the various carbonates studied. This indicates a similar energy barrier, or, in other words, the same kinetic process is operative for all the liquids.

## 2. Dielectric Relaxation

From the relaxation frequencies  $f_R$  one may calculate relaxation times  $\tau^{-1} = 2\pi f_R$ , which in turn can be converted into microscopic relaxation times  $\tau_m$  by the Powless-Glarum expression<sup>12</sup> (among other options):

$$\tau_m = \frac{2\epsilon_0 + \epsilon_\infty}{3\epsilon_0} \tau$$

$\tau_m$  may be related to a frictional coefficient (which is associated to hydrodynamic frictional forces) by the equation  $\tau_m = \xi / 2kT$ . For a spherical dipole of radius  $a$ , rotating in a medium with bulk viscosity  $\eta$ ,  $\xi = 8\pi a^3 \eta$  (see Ref. 7) or

$$\tau_m = \frac{4\pi a^3}{kT} \eta = \frac{3v}{kT} \eta$$

where  $v$  is the volume of the spherical dipole.

The values of  $\tau_m$  and  $\eta$  found in Table II have been plotted in the form of  $\tau_m$  versus  $\eta$  in Figure 7. If the volume of the rotating dipolar specie is constant for the series of non-cyclic dialkylcarbonate studied, then the plot should be linear with slope =  $3v/kT$ . A roughly linear plot is obtained by linear regression (correlation coefficient  $r^2 = 0.987$ ). This corresponds to a rotating volume  $v = 1.48_4 \times 10^{-23} \text{ cm}^3$  or  $\bar{V} = vL = 8.93 \text{ cm}^3/\text{mole}$  ( $L$ , Avogadro's number). These numbers are about one order of magnitude smaller than the molar or molecular volume of the carbonates as computed from the expressions:  $\bar{V} = M/\rho$ ,  $v = M/L\rho$ , where  $M$  is the molecular weight (in grams/mole) and  $\rho$  the density of the corresponding carbonate (Table II).

One would then be tempted to conclude from the above, that  $\tau_m$  does not refer to molecular tumbling of the whole carbonate molecule but only of a part of it. Similar conclusions have been reached by authors<sup>13</sup> dealing with the interpretation of the dielectric relaxation of other compounds containing the alkoxy group, namely, the esters.

On the other hand, it has been noted<sup>14</sup> that the application of the Debye hydrodynamic model of dipole reorientation leads to unreasonably small values of the molecular radius. Modification of the Debye relationship by multiplying the bulk viscosity by a factor  $f$  of the order of 0.1-0.2, according to Gierer and Wirtz<sup>15</sup> gives radii of the expected order of magnitude. Rotational correlation times  $\tau_\theta$  from NMR spectra have

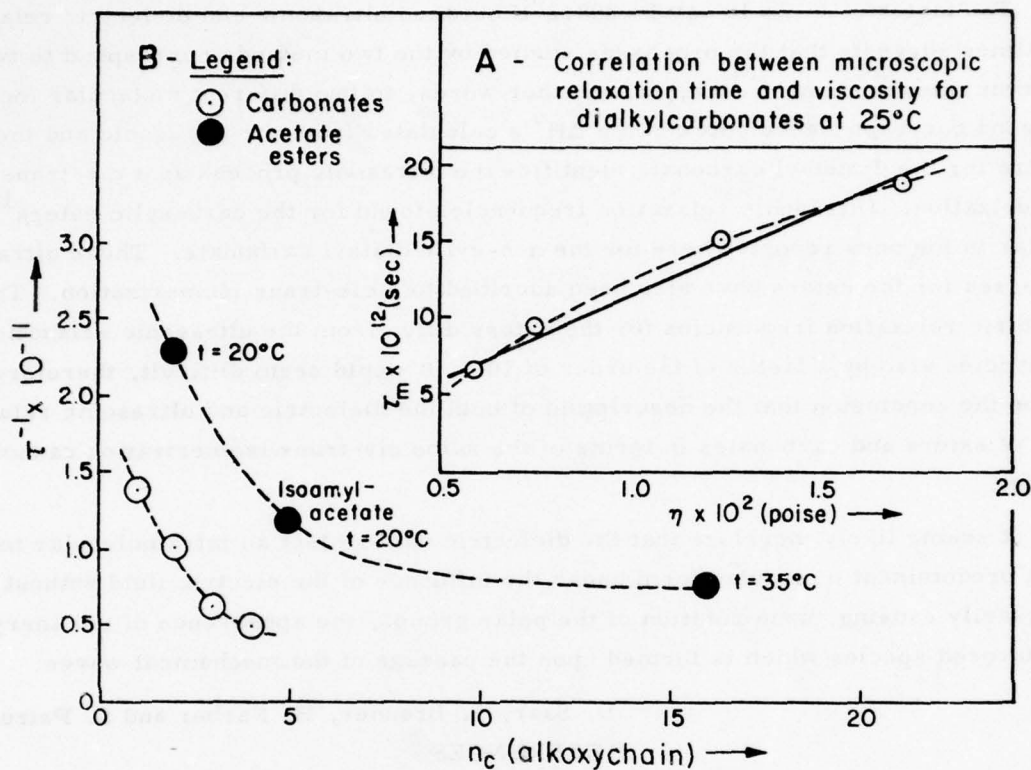


Fig. 7. Correlation between the inverse of the dielectric relaxation times  $\tau^{-1}$  and the number of carbon atoms of the alkoxy chains for liquid carbonates and esters. Inset: Molecular dielectric relaxation time  $\tau_m$  vs. the viscosity  $\eta$  for the liquid carbonates investigated at 25°C.

also been correlated to the bulk viscosity by McClung and Kivelson<sup>16</sup> with the introduction of a similar factor.

There is, however, some further evidence from Fig. 7 which would seem to favor the interpretation of the relaxation process as segmental motion rather than of whole molecule tumbling.

In Fig. 7 is shown the plot of the reciprocal of the dielectric relaxation time  $\tau^{-1}$  versus the number of carbon atoms in the alkyl chain of the alkoxy group. One sees a decreasing change in  $\tau^{-1}$  as the number of carbon atoms is increased. This is more evident for the esters, given the much higher chain lengths investigated.<sup>13</sup> This behavior suggests that the observed relaxation times refer to a segmental motion of the molecule, probably involving the alkoxy groups. However, by elongating the latter over a certain number of carbon atoms, the relaxation times tend to become constant due to nonparticipation in the motion of carbon atoms of the alkyl chain far away from the rotating groups.<sup>13</sup>

### 3. Comparison Between Ultrasonic and Dielectric Relaxation Processes

The factor of  $1 - 2 \times 10^3$  at  $T = 298.2^\circ\text{K}$  between ultrasonic and dielectric relaxation times suggests that the processes studied by the two methods correspond to two different rate constants  $\tau^{-1} = k$ , or, in other words, to two different molecular motions. The good correspondence between the  $\Delta H^\circ$ 's calculated from the ultrasonic and the IR spectra for the dimethyl carbonate identifies the ultrasonic process as a cis-trans isomerization. Ultrasonic relaxation frequencies found for the carboxylic esters<sup>17</sup> are similar to the ones reported here for the non-cyclic dialkyl carbonate. These ultrasonic processes for the esters have also been ascribed to a cis-trans isomerization. The dielectric relaxation frequencies for the esters differ from the ultrasonic relaxation frequencies also by a factor of the order of  $10^3$ . It would seem difficult, therefore, to escape the conclusion that the description of both the dielectric and ultrasonic relaxations of esters and carbonates in terms of the same cis-trans isomerization cannot be correct.

It seems likely therefore that the dielectric data reflect an intramolecular motion of the predominant molecular form under the influence of the electric field without necessarily causing, upon rotation of the polar groups, the appearance of the energetically unfavored species which is formed upon the passage of the mechanical waves.

D. Saar, J. Brauner, H. Farber and S. Petrucci

#### REFERENCES

1. J. Lamb in "Physical Acoustics," Vol. II, part A, Chap. 4, Ed. W.P. Mason (New York: Academic Press, 1965).
2. C.P. Smyth, "Dielectric Behavior and Structure," (New York: McGraw Hill, 1955).
3. S. Petrucci in "Ionic Interactions," (New York: Academic Press, 1971) Vol. II, Ch. II.
4. M. Eigen and L. DeMaeyer in "Rates and Mechanisms of Reactions," part II, A. Weissberger, Ed., Ch. XVIII, Interscience 2nd edition (1963).
5. A. Fanelli and S. Petrucci, J. Phys. Chem. 75, 2649 (1971); S. Petrucci and M. Battistini, J. Phys. Chem. 71, 1181 (1967); S. Petrucci, J. Phys. Chem. 81 (1967).
6. D. Saar, J. Brauner, H. Farber and S. Petrucci, J. Phys. Chem. 82, 545 (1978); H. Farber and S. Petrucci, J. Phys. Chem. 79, 1221 (1975).
7. N. Hill in "Dielectric Properties and Molecular Behavior," (London: Van Nostrand, 1969) Ch. I.
8. M.W. Evans, M.N. Afsar, G.J. Davies, C. Menard and J. Goulon, Chem. Phys. Lett. 52, 388 (1977).
9. Ref. 1, p. 241.
10. J.E. Katon and M.D. Cohen, Can. J. Chem. 52, 1994 (1974).
11. H. Farber and S. Petrucci, J. Phys. Chem. 80, 327 (1976).
12. M. Davies in "Dielectric Properties and Molecular Behavior," (London: Van Nostrand Co., 1969) p. 298; J.G. Powles, J. Chem. Phys. 21, 633 (1953); S.H. Glarum, J. Chem. Phys. 33, 1371 (1960); R.H. Cole, J. Chem. Phys. 42, 637 (1965).

13. a) Ref. 2, pp. 121-123; b) P.L. McGeer, A.J. Curtis, G.B. Rathmann and C.P. Smyth, J. Am. Chem. Soc. 74, 3541 (1951); c) see also Y. Koga, H. Takahashi, K. Higasi, Bull. Chem. Soc. Japan 47, 84 (1974) for solutions of alkyl esters in benzene.
14. R. Payne and I.E. Theodorov, J. Phys. Chem. 76, 2892 (1972).
15. A. Gierer and K. Wirtz, Z. Naturforsch. 8a, 532 (1953).
16. R.E.D. McClung and D. Kivelson, J. Chem. Phys. 49, 3380 (1968).
17. J. Bailey and A.M. North, Trans. Faraday Soc. 64, 1497 (1968); J. Karpovich, J. Chem. Phys. 22, 1767 (1954).

## DIELECTRIC RELAXATION OF SOME 1:1 ELECTROLYTES IN TETRAHYDROFURAN AND DIETHYLCARBONATE

D. Saar, J. Brauner, H. Farber, S. Petrucci

Complex permittivities in the frequency range of 0.3 - 67 GHz are reported for the systems  $\text{LiNO}_3$  0.1M, Na-picrate 0.05M,  $\text{Bu}_4\text{NNO}_3$  0.1M in THF at 25°C and for the systems  $\text{LiClO}_4$  0.1M, and  $\text{LiSCN}$  0.1M in diethylcarbonate (DEC) at 25°C. The permittivities of the electrolyte solutions can be described within the experimental error, by the sum of two Debye relaxation processes. The high frequency process is comparable to the one observed for the pure solvents while the low frequency process is due to the presence of the electrolytes.

Apparent dipole radii  $\underline{a}_\tau$  for the electrolytes in the THF solutions are calculated from the Debye relation using the macroscopic viscosity of the solvent. Charge to charge separation distances  $\underline{a}_\mu$  are calculated from the Böttcher relation which relates the change in dielectric strength ( $\epsilon_0 - \epsilon_\infty$ ), for a dipolar relaxation process, with the apparent dipole moment. These two parameters  $\underline{a}_\tau$  and  $\underline{a}_\mu$  determined for the present systems and two previous systems ( $\text{LiClO}_4$  and  $\text{NaClO}_4$  in THF) are linearly related to each other (correlation coefficient  $r^2 = 0.89$ ). The  $\underline{a}_\tau$ 's also correlate linearly ( $r^2 = 0.91$ ) with the sums of the crystallographic radii. These relations reinforce our previous conclusions that the solute relaxation process in THF is mainly due to the rotation of ion-pair dipoles.

For the DEC solutions the appearance of a single relaxation process for the solute (instead of a distribution of relaxation times as the presence of two molecular species would suggest) is discussed in terms of the structure of the quadrupoles. Recent Raman and IR spectra in the literature suggest a centrosymmetric structure for  $(\text{LiSCN})_2$  in ethers. This structure if present for the carbonate solutions would predict the absence of a quadrupole dielectric relaxation process in accord with our results in dimethylcarbonate (DMC) and DEC.

#### A. Introduction

In the past, complex permittivities of electrolytes in solvents of low permittivity have been measured in order to characterize the dynamic behavior of the electrolytes under the influence of microwave electric fields and to study the relaxation behavior of the solute molecular species. The general picture which had emerged from the previous studies carried out in this laboratory<sup>1,2,3</sup> was that the solute dielectric relaxation process was mainly caused by the rotational relaxation of the ion-pair dipoles. These species were also present in the largest concentration. However, these studies were confined to some alkali perchlorates in THF (Refs. 1, 2) and to lithium salts in DMC (Reference 3.)

It was of interest to try to generalize the above conclusions for electrolytes of different nature and structure in the same solvent. Therefore, we have now extended our studies to include  $\text{LiNO}_3$ , Na-Picrate, and  $\text{Bu}_4\text{NNO}_3$  in THF in the same concentration range as used previously.

The study of  $\text{Li}^+$  salts in DMC had shown a correlation between the change in the dielectric strength of the system  $\epsilon_0 - \epsilon_{\infty 1}$  due to the solute relaxation process and the calculated concentration of ion-pair dipoles,  $C_p$ . It was of interest to extend this work to another solvent, namely DEC. It has been shown<sup>4</sup> that although DEC is practically isodielectric with DMC,  $\text{Li}^+$  salts form ion-pair dimers or quadrupoles to a much larger extent in DEC. To this end we have studied  $\text{LiClO}_4$  (0.1M) and  $\text{LiSCN}$  (0.1M) in DEC by dielectric relaxation methods to ascertain whether the structural information inferred by Raman and IR spectra is reflected by the molecular dynamics of the systems.

### B. Experimental Part

The equipment and procedure have been described elsewhere.<sup>1,3</sup> THF (Fisher) was distilled under nitrogen over liquid potassium with only the middle portion collected. DEC (Aldrich) was distilled under reduced pressure. The  $\text{LiNO}_3$  (Fisher) was dried in an oven at  $101^\circ\text{C}$  to constant weight. The NaPicrate (Kodak) and  $\text{Bu}_4\text{NNO}_3$  (Kodak) were dried under  $\sim 0.1$  torr at  $30^\circ$  for one week. The  $\text{LiClO}_4$  and  $\text{LiSCN}$  were purified as described previously.<sup>3</sup>

### C. Results

Figures 1 and 2 show the real part of the complex permittivity  $\epsilon'$  and the imaginary part of the complex permittivity  $\epsilon''$  plotted versus frequency  $f$ . Both figures include a Cole-Cole plot of  $\epsilon'' - \epsilon''_x$  vs.  $\epsilon'$ , where  $\epsilon''_x$  represents the (dielectric) loss due to ionic conduction. The data are for the systems  $\text{LiNO}_3$  (0.1M), NaPicrate (0.05M) and  $\text{Bu}_4\text{NNO}_3$  (0.1M) in THF at  $25^\circ\text{C}$ . The filled in points at  $f = 137$  GHz are literature data for pure THF (Reference 5.)

In both figures the solid lines represent the functions  $\epsilon'$  and  $\epsilon''$  vs.  $f$  as calculated from the sums of two Debye functions:

$$\epsilon' = \epsilon_{\infty 2} + (\epsilon_0 - \epsilon_{\infty 1}) / [1 + (f/f_{R1})^2] + (\epsilon_{\infty 1} - \epsilon_{\infty 2}) / [1 + (f/f_{R2})^2] \quad (1)$$

$$\epsilon'' - \epsilon''_x = (\epsilon_0 - \epsilon_{\infty 1})(f/f_{R1}) / [1 + (f/f_{R1})^2] + (\epsilon_{\infty 1} - \epsilon_{\infty 2})(f/f_{R2}) / [1 + (f/f_{R2})^2] \quad (2)$$

In the above  $\epsilon_0$  is the static permittivity,  $\epsilon_{\infty 1}$  is the permittivity at  $(f \gg f_{R1})$ ,  $\epsilon_{\infty 2}$  is



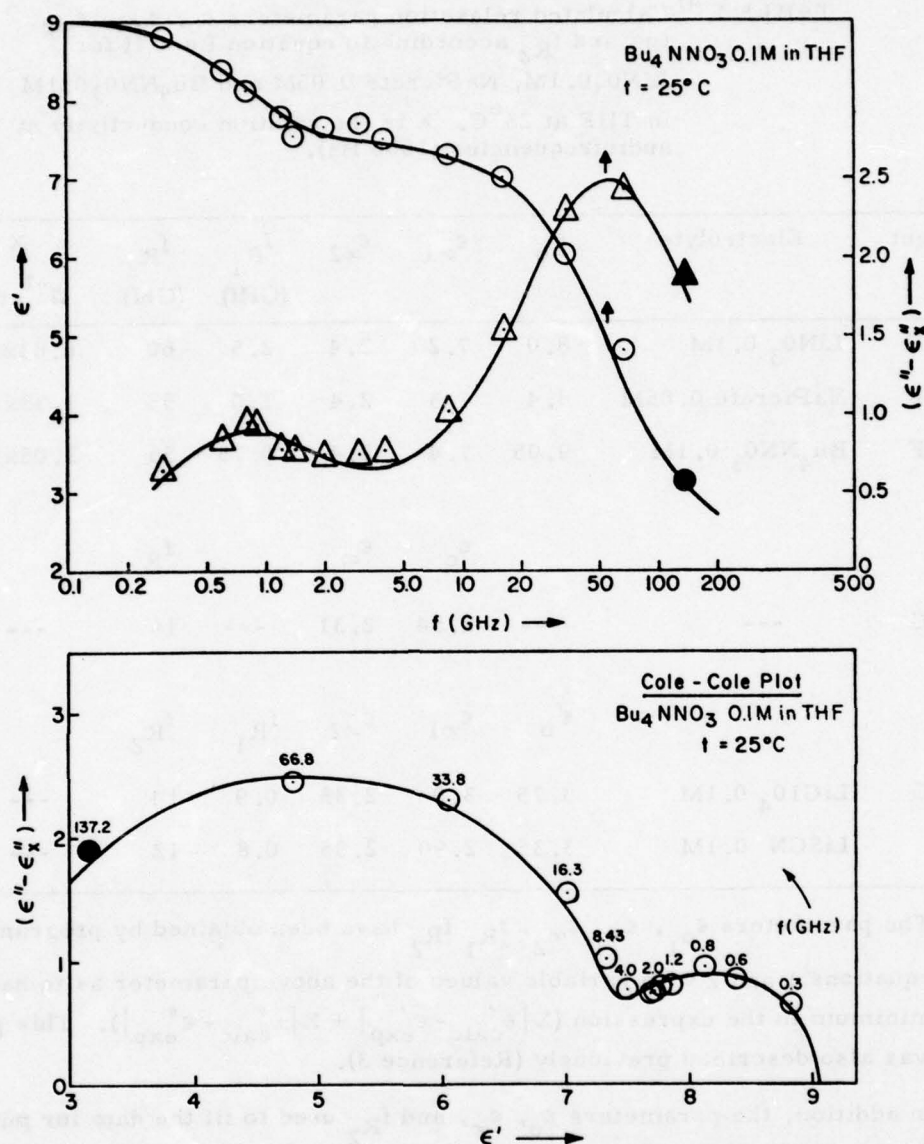


Fig. 2. Real part of the complex permittivity  $\epsilon'$  and imaginary part of the complex permittivity  $\epsilon'' - \epsilon''_x$  plotted vs. the frequency  $f(\text{GHz})$  for  $\text{Bu}_4\text{NNO}_3$  0.1M in THF at  $25^\circ\text{C}$ . Cole-Cole plot of  $\epsilon'' - \epsilon''_x$  vs.  $\epsilon'$  for the same system.

The filled points of  $f=137$  GHz refer to literature data for pure THF (Ref. 4).

the high frequency permittivity ( $f \gg f_{R2}$ ), and  $f_{R1}$  and  $f_{R2}$  are the relaxation frequencies. The parameters  $\epsilon_0, \epsilon_{\infty 1}, \epsilon_{\infty 2}, f_{R1}$  and  $f_{R2}$  are given in Table I, together with the fitting procedure.

$\epsilon''_x$  is calculated from the expression  $\epsilon''_x = 1.8 \times 10^{12} X/f$  where the conductivity of the solutions,  $X$ , was measured separately at  $25.00^\circ\text{C}$ . The conductivity apparatus consisted of a General Radio impedance comparator (equipped with an impedance reference arm), a Kraus conductance cell, and a precision thermostat as described earlier.<sup>5</sup> The values of  $X$  are also given in Table I.

TABLE I. <sup>a)</sup> Calculated relaxation parameters  $\epsilon_0$ ,  $\epsilon_{\infty 1}$ ,  $\epsilon_{\infty 2}$ ,  $f_{R1}$  and  $f_{R2}$  according to equation I and II for  $\text{LiNO}_3$  0.1M, NaPicrate 0.05M and  $\text{Bu}_4\text{NNO}_3$  0.1M in THF at 25°C. X is the solution conductivity at audiofrequencies (1000 Hz).

Solvent	Electrolyte	$\epsilon_0$	$\epsilon_{\infty 1}$	$\epsilon_{\infty 2}$	$f_{R1}$ (GHz)	$f_{R2}$ (GHz)	X $\Omega^{-1}\text{cm}^{-1}$
THF	$\text{LiNO}_3$ 0.1M	8.0	7.2	2.4	2.5	60	$6.83 \times 10^{-6}$
THF	NaPicrate 0.05M	8.4	7.3	2.4	1.0	55	$1.33 \times 10^{-5}$
THF	$\text{Bu}_4\text{NNO}_3$ 0.1M	9.05	7.4	2.4	0.75	55	$2.05 \times 10^{-4}$
			$\epsilon_0$	$\epsilon_{\infty}$		$f_R$	
DEC	---	---	2.84	2.31	---	16	---
		$\epsilon_0$	$\epsilon_{\infty 1}$	$\epsilon_{\infty 2}$	$f_{R1}$	$f_{R2}$	
DEC	$\text{LiClO}_4$ 0.1M	3.75	3.0	2.35	0.9	14	---
	LiSCN 0.1M	3.35	2.90	2.35	0.8	12	---

a) The parameters  $\epsilon_{\infty 1}$ ,  $\epsilon_{\infty 2}$ ,  $f_{R1}$ ,  $f_{R2}$  have been obtained by programming equations 1 and 2 with variable values of the above parameter as to have a minimum in the expression  $(\sum |\epsilon'_{\text{calc}} - \epsilon'_{\text{exp}}| + \sum |\epsilon''_{\text{calc}} - \epsilon''_{\text{exp}}|)$ . This procedure was also described previously (Reference 3).

In addition, the parameters  $\epsilon_0$ ,  $\epsilon_{\infty}$ , and  $f_{R2}$  used to fit the data for pure DEC via one term Debye relaxation functions, (Eqs. (3) and (4)) are presented in Table I,

$$\epsilon' = \epsilon_{\infty} + (\epsilon_0 - \epsilon_{\infty}) / [1 + (f/f_R)^2] \quad (3)$$

$$\epsilon'' = (\epsilon_0 - \epsilon_{\infty}) (f/f_R) / [1 + (f/f_R)^2] \quad (4)$$

The fitting of the DEC solvent by a single relaxation function is adequate up to about 35 GHz with positive deviations appearing at high frequency due to an absorption process in the submillimeter near infrared region of wavelengths. Details of this aspect will

be described in a later paper. For the time being, our interest being centered on solute relaxation, we will ignore the systematic deviations of  $\epsilon''$  at 50 - 70 GHz.

Figure 3 shows the quantities  $\epsilon'$  and  $\epsilon''$  versus the frequency  $f$  and  $\epsilon''$  versus  $\epsilon'$  for  $\text{LiClO}_4$  (0.1M) and  $\text{LiSCN}$  (0.1M) in DEC at 25°C. The solid lines correspond to values calculated from Eqs. (1) and (2) where  $\epsilon''_x$  has been neglected. The associated parameters are reported in Table I.

#### D. Discussion

Literature<sup>5,7</sup> reports assigns a single Debye type relaxation process to pure THF with parameters  $\epsilon_0 = 7.36$ ,  $\epsilon_\infty = 2.30$ , and  $f_R = 62.9$  GHz. From Figs. 1 and 2, and Table I it is evident that the high frequency relaxation process in the THF solutions can be associated with the solvent molecules. The second relaxation process occurring at lower frequencies in the solutions is specific, both in terms of the change in dielectric strength  $\epsilon_0 - \epsilon_\infty$  and relaxation frequency  $f_{R1}$ , for the electrolyte considered.

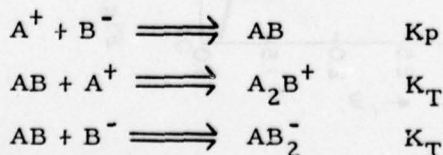
In the following we will relate apparent dipolar radii obtained from the relaxation parameters  $\epsilon_0 - \epsilon_\infty$  and from the relaxation time,  $\tau$ , [where  $\tau = (2\pi f_R)^{-1}$ ], to each other and to radii calculated from crystallographic data. These relationships should clarify the nature of the molecular process associated with the dielectric relaxation attributed to the solute.

First, we present equations which will permit the calculation of  $a_\mu$ , the apparent distance between the centers of unit positive and negative charge for the ion-pair dipole where  $a_\mu = \mu/e$ ,  $\mu$  is the apparent dipole moment of the ion-pair and  $e$  is the electron charge. The apparent dipole moment  $\mu$  is calculable from the Böttcher expression<sup>8</sup>

$$\epsilon_0 - \epsilon_\infty = \frac{4\pi LCp \times 10^{-3} \epsilon_0 \mu^2}{(1 - \alpha f)^2 (2\epsilon_0 + 1) 3kT} \quad (5)$$

where  $Cp$  is the concentration of the ion-pair dipoles,  $L$  is Avogadro's number,  $f$  is the reaction field factor, and  $\alpha$  the polarizability. To use Eq. (5) we need to calculate  $CP$  which can be done from values of the equilibrium constants for ion-pair formation and triple-ion formation,  $K_p$  and  $K_T$ , respectively. The values  $K_p$  and  $K_T$  are in the literature<sup>9</sup> and are presented in Table II.

The equilibria involved are:



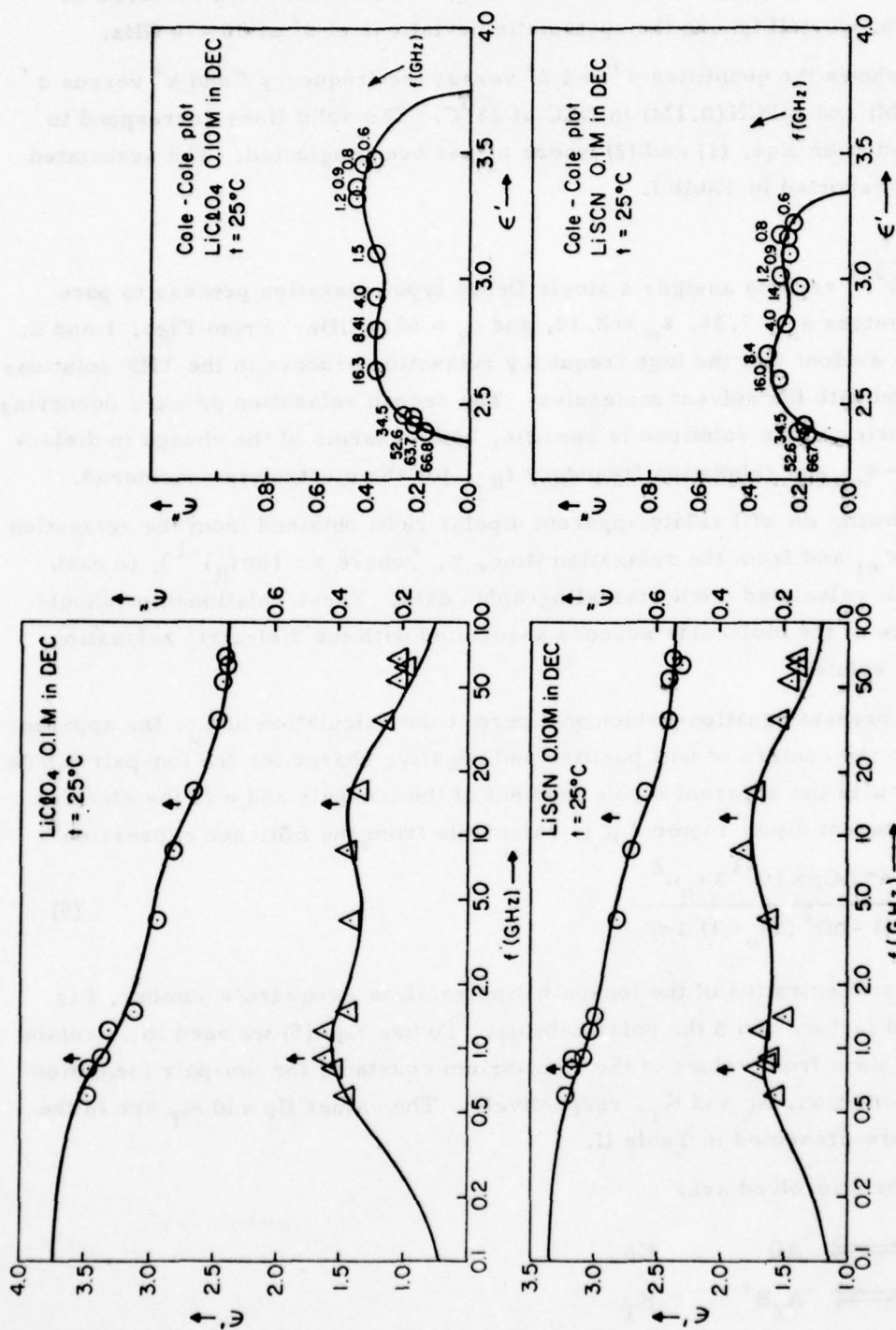


Fig. 3. Real part of the complex permittivity  $\epsilon'$  and imaginary part of the complex permittivity  $\epsilon''$  plotted vs. the frequency  $f$  (GHz) for  $\text{LiClO}_4$  0.1M and  $\text{LiSCN}$  0.1M in DEC at  $25^\circ\text{C}$ . Cole-Cole plots of  $\epsilon''$  vs.  $\epsilon'$  for the same systems.

TABLE II. Values of ion pairs and triple ions formation constants  $K_p$  and  $K_T$  in the THF taken from the literature; concentration of ion pairs  $C_p = (1 - \sigma - 3\sigma_T)C$ ; charge to charge separation  $a_u$  calculated according to Debye; sum of the crystallographic radii ( $r_+ + r_-$ ) taken from the literature and from molecular models.

Electrolyte in THF	C M	$K_p$ $M^{-1}$	$K_T$ $M^{-1}$	Ref.	$C_p$ M	$a_u 10^8$ cm	$a_T \times 10^8$ ( $r_+ + r_-$ ) $10^8$	Ref.
$LiClO_4$	0.05	$4.84 \times 10^7$	153	a)	0.0492	3.11	4.18	2.25 d)e)
$NaClO_4$	0.1	$9.93 \times 10^7$	150	b)	0.0985	2.63	3.78	2.62 e)
$NaClO_4$	0.05	$9.93 \times 10^7$	150	b)	0.0495	3.02	3.93	2.62 e)
$LiNO_3$	0.10	$590 \times 10^7$	132	c)	0.0993	2.00	3.53	1.82 f)
NaPicrate	0.05	---	---	-	(0.5)	4.36	4.77	4.5 g)
$Bu_4NNO_3$	0.10	$0.7 \times 10^7$	155	c)	0.0933	4.10	5.22	6.16 h)

#### References

- P. Jagodzinski and S. Petrucci, J. Phys. Chem. 78, 917 (1974).
- H. Farber and S. Petrucci, J. Phys. Chem. 80, 327 (1976).
- H. C. Wang and P. Hemmes, J. Am. Chem. Soc. 95, 5119 (1973).
- 0.60 Å is the radius of the  $Li^+$  ion.
- 0.97 Å is the radius of the  $Na^+$  ion; 1.65 Å is the Cl-O distance in the perchlorate ion.
- 1.22 Å is the N-O distance in the nitrate ion.
- 3.5 Å is the calculated radius of the picrate ion from molecular models (distance from the oxygen to the center of the benzene ring).
- $r_{Bu_4N^+} = 4.94$  Å, from R. A. Robinson and R. H. Stokes, "Electrolyte Solutions," Butterworths, 2nd Ed., page 125 (1959).

Now  $K_p = (1 - \sigma)/\sigma^2 C$  where  $\sigma$  is the degree of dissociation of the pairs and C is the stoichiometric concentration of the electrolyte. Similarly,<sup>10</sup>

$$K_T = \frac{[AB_2]}{[AB][A]} = \frac{[A_2B]}{[AB][B]} = \frac{\sigma_T C}{[(1 - \sigma - 3\sigma_T)C](\sigma C)}$$

where  $\sigma_T$  is the fraction of ion-pairs which form triple ions according to the equilibria for triple ion formation shown above. Obviously then,  $C_p = [AB] = (1 - \sigma - 3\sigma_T)C$ .

The  $C_p$ 's were calculated for all the THF solutions except the Na-picrate solution where no data for  $K_p$  and  $K_T$  are available. In the latter case  $C_p$  was assumed equal to  $C$ , the total solute concentration. These values are presented in Table II.

The  $(1 - \alpha f)^2$  factor in Eq. (5) was neglected which implies a zero polarizability or a rigid sphere model for the ions in the ion-pair dipole. Thus in the rigid sphere model for the constituent ions  $a_\mu$  is equal to  $a_\mu = \frac{\mu}{e}$ , the charge to charge separation distance. The values of the  $\mu$ 's and  $a_\mu$ 's are reported in Table II.

As an alternate method for the calculation of an apparent radius for the relaxing dipole we use Debye's equation

$$\tau' = \frac{4\pi a_\tau^3}{kT} \eta \quad (6)$$

where  $a_\tau$  is the radius of a spherical dipole immersed in a continuum of viscosity  $\eta$  and  $\tau'$  is the microscopic relaxation time. This can be related to the experimental relaxation time  $\tau$  by the Powless-Glarum expression<sup>11</sup>  $\tau' = [(2\epsilon_0 + \epsilon_\infty)/3\epsilon_0] \tau$ . The values of  $a_\tau$  calculated from Eq. (6) are given in Table II.

Figure 4 shows the values of  $a_\tau$  versus  $a_\mu$ . The solid straight line is that calculated by a linear regression analysis with a correlation coefficient  $r^2 = 0.89$ . We would not

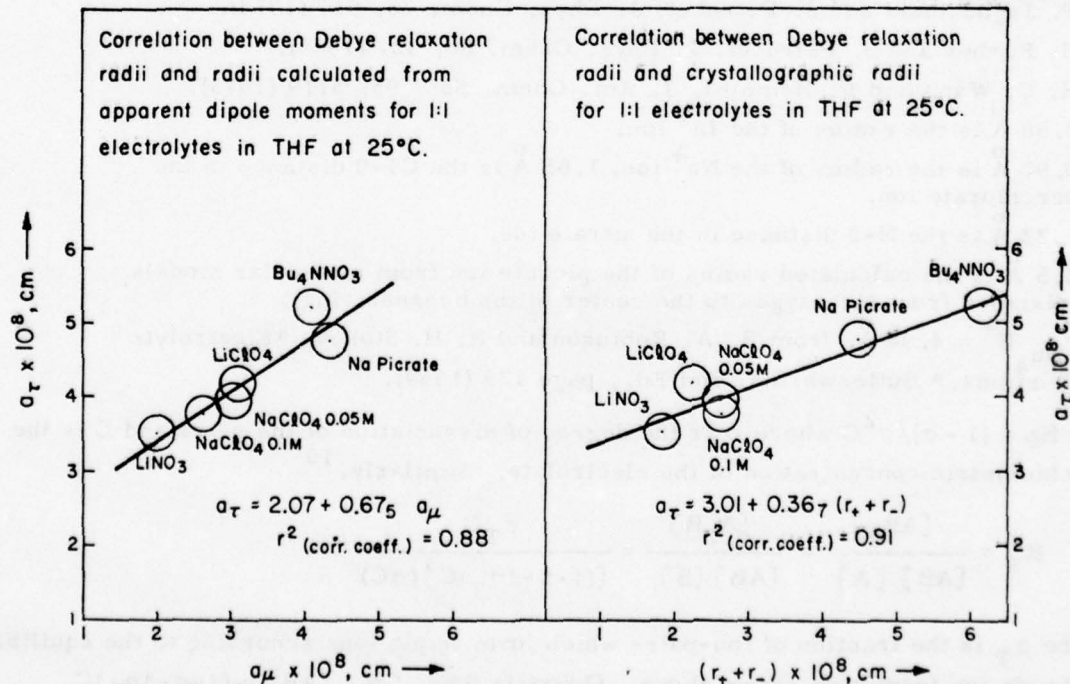


Fig. 4. Correlation between  $a_\tau$ , the dipole radius calculated from the relaxation time and  $a_\mu$  the charge to charge distance calculated from the apparent dipole moment. Correlation between  $a_\tau$  and the sum of the crystallographic radii ( $r_+ + r_-$ ).

expect a slope of one in this plot in view of the crudeness of the models assumed and the fact that  $a_{\mu}$  is derived from a property related to the distribution of charge in the ion-pair while  $a_{\tau}$  is derived from a property related to the rotation of the ion-pair in the solution. The ion-pair may be solvated to some extent and this would modify the size of the rotating species. In spite of all this a correlation as shown in Fig. 4 does emerge.

Further evidence that the solute relaxation process is due to the rotational relaxation of ion-pair dipoles is given by the linear relationship between the  $a_{\tau}$ 's and the sums of the crystallographic radii ( $r_{+} + r_{-}$ ) also shown in Figure 4. The solid straight line was calculated from a linear regression analysis with a correlation coefficient of  $r^2 = 0.91$ .

Attempts to rationalize the observed correlation between  $a_{\tau}$  and the sum of the crystallographic radii in terms of a libration of an ionic lattice in the alternating electric field of the microwaves, while not impossible, is not likely to be justified. Explanations of this sort would be more reasonable at higher concentrations than those studied, where the pair to pair distances would become the same as the interionic distances in a single pair. At the maximum concentration studied here,  $C = 0.1M$ , (neglecting other species) the volume associated with a single pair,  $V = 1/(Cx10^{-3}L) = 1.66 \times 10^{-20} \text{ cm}^3$ . Taking  $V = \frac{4}{3} \pi r^3$  this volume corresponds to  $\approx 16 \text{ \AA}$  where  $2r$  represents the minimum distance between the centers of a given pair and some adjacent pair. The value of the calculated  $2r$  is between 7 and 16 times the values of  $a_{\mu}$  quoted in Table II.

In a previous paper<sup>3</sup> we reported a roughly linear correlation between the quantity  $\epsilon_0 - \epsilon_{\infty 1}$  and  $Cp \ 3\epsilon_0/2\epsilon_0 + 1$  for LiBr, LiSCN, and LiClO<sub>4</sub> in DMC in accord with Equation (5). In that correlation the contribution to  $\epsilon_0 - \epsilon_{\infty 1}$  due to rotational relaxation of any ion-pair dimers (quadrupoles) was neglected because of the small dipole moments of these dimers for LiBr and because of their low concentration for LiClO<sub>4</sub> (Reference 4).

For the present systems of LiSCN and LiClO<sub>4</sub> in DEC, the association to quadrupoles should be more extensive and the dipole moments of these quadrupoles should be relatively large,<sup>4</sup> as indicated by the values of  $\delta_q$ . The  $\delta_q$ 's represent the contribution per unit concentrations of the quadrupoles to the increase of the permittivity of the solution with respect to the solvent,  $\Delta\epsilon$ . Then it would appear that the approximation of neglecting the contribution of the quadrupoles in the quantity  $\epsilon_0 - \epsilon_{\infty}$  is no longer tenable for the present systems.

Still it is not possible to detect the presence of more than one single Debye relaxation process, within experimental error, for 0.1M LiSCN and LiClO<sub>4</sub> in DEC. It is

not conceivable that the two species ion-pair dipoles and quadrupole-dimers have the same relaxation time. On the other hand, the presence of the latter species has been detected by IR spectra ( $\sim 2040 \text{ cm}^{-1}$  band).<sup>13</sup>

Since the parameter  $\epsilon_{\infty 2}$  is comparable for the electrolyte solution and the solvent DEC (and the size of the quadrupoles is larger than the one of the dipoles), one might suspect that the rotational relaxation of these species occur at lower frequency than our lowest experimental frequency. That this suspicion is unfounded is easily provable by recalculating the static  $\epsilon_0$ 's (at  $f = 2 \text{ MHz}$ ) from Chabanel's work<sup>4</sup> and comparing it with our calculated  $\epsilon_0$ 's.

Menard and Chabanel<sup>4</sup> defined the dielectric effect  $\Delta\epsilon$

$$\Delta\epsilon = \delta_p C_p + \delta_q C_q \quad (7)$$

where  $\Delta\epsilon = \epsilon_0 - \epsilon_s$  with  $\epsilon_s$  the permittivity of the solvent and  $C_p$  and  $C_q$  are the concentration of the pairs and quadrupoles respectively.  $\delta_p$  and  $\delta_q$  are the dielectric increments due to the presence of unit concentrations of the two species.

The calculation of  $\Delta\epsilon$  is possible from the tabulated<sup>4</sup>  $\delta_p$ ,  $\delta_q$  and  $K_q$  (Ref. 4), where  $K_q$  refers to the equilibrium:



$$K_q = \frac{\sigma_q/2}{C(1 - \sigma_q)^2}$$

$$\text{and } C_p = (1 - \sigma_q)C, \quad C_q = \frac{\sigma_q}{2} C$$

Table III reports the results of this calculation for LiSCN and LiClO<sub>4</sub> in DEC as well as for LiBr, LiSCN in DMC (the values for  $K_q$  and  $\delta_p$  and  $\delta_q$  for LiClO<sub>4</sub> in DMC were not reported.<sup>4</sup> It is evident that the two sets of  $\epsilon_0$ 's, the static and the extrapolated ones are within experimental error.

The absence of a dielectric relaxation spectrum for the quadrupoles can be rationalized by the possibility that dipole moment is negligible or zero at variance with calculated values.<sup>4</sup> Recently, Chabanel et al.<sup>14</sup> have obtained the IR and Raman spectra of LiSCN in various ethers. They invariably detect a lowering of the force constant in the C-N bond of the SCN<sup>-</sup> ion when the ion-pairs dimerize to form

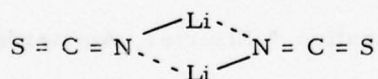
TABLE III. Calculation of  $\Delta\epsilon = \epsilon_o - \epsilon_s = \delta_p C_p + \delta_q C_q$  for LiSCN 0.1M, LiClO<sub>4</sub> 0.1M in DEC and for LiBr 0.1M and LiSCN 0.09M in DMC.  $\epsilon_s$  is the solvent permittivity. Comparison between  $\epsilon_o$  at 2MHz from  $\Delta\epsilon$  and  $\epsilon_o$  from the microwave data

Solvent	Electrolyte	$C_p$	$\delta_p$	$C_q$	$\delta_q$	$\Delta\epsilon$	$\epsilon_o$ (a)	$\epsilon_o$ <sup>extrap</sup> (b)
DEC	0.1M LiSCN	0.017	14.3	0.0414	8.0	0.577	3.42	3.35
DEC	0.1M LiClO <sub>4</sub>	0.021	14.7	0.0395	15.0	0.902	3.74	3.75
DEC	0.1M LiBr	0.021	8.7	0.0395	1.8	0.254	3.34	3.45
DEC	0.09M LiSCN	0.037	14.6	0.0260	10.8	0.830	3.92	3.95

(a)  $\epsilon_o = \Delta\epsilon + \epsilon_s$ ,  $\epsilon_s$  (DEC) = 2.84;  $\epsilon_s$  (DMC) = 3.09

(b) Data from Table I and Reference 3.

quadrupoles. In fact, the "free" SCN antisymmetric stretching frequency at 2060 cm<sup>-1</sup> is lowered to 2040 - 2030 cm<sup>-1</sup>. Similarly, the IR active C-S stretching frequency is increased from 735 cm<sup>-1</sup> in the "free" ion (which occurs in solvents such as dimethylformamide) to 792 cm<sup>-1</sup> on quadrupole formation in dimethoxymethane. In contrast, the formation of the simple ion-pair LiSCN results in increases in both the C-N and C-S stretching frequencies with respect to the free SCN<sup>-</sup> ion. The decrease of the force constant in the C-N bond and the simultaneous increase in that of the C-S bond had been interpreted<sup>14</sup> as being due to the formation of a centrosymmetric quadrupole



By symmetry the dipole moment of this species would be zero.

Unfortunately, the carbonate solvents absorb in the same region as the C-S stretch in the thiocyanate ion which precludes the study of the thiocyanate-carbonate system in this region. However, for the C-N stretching frequency IR data<sup>13</sup> indicate the presence of quadrupoles in carbonate solvents, the band occurring at 2040 cm<sup>-1</sup>. (Raman spectra have confirmed this band although at 2052 cm<sup>-1</sup> (Reference 3)). Addition of a high permittivity solvent such as dimethylformamide to the LiSCN-ethylacetate system also causes this IR band at 2040 cm<sup>-1</sup> to disappear.<sup>13</sup>

If the ion-pair dimers or quadrupoles that are formed in the carbonate solvents are centrosymmetric and thus have a zero dipole moment as is indicated in the case

of the ether solvents, then the dielectric data for the LiSCN and LiClO<sub>4</sub> in DEC would find a complete rationalization for their showing a single relaxation process. This then is to be attributed to the ion-pair dipoles as in the previous case for the DMC solvent.<sup>4</sup>

According to the report of one of the referees<sup>1</sup>, Chabanel et al have recently arrived at the same conclusion. The apparent inconsistency between IR( $\mu_q = 0$ ) Eq. (14) and dielectric ( $\delta q \neq 0$ ) measurements Eq. (4) was puzzling. A quantitative IR study of LiSCN solutions in DEC and in other solvents showed that the activity coefficients of ion pairs cannot be neglected in the 0.01 - 0.1M concentration range. Thus the ideal solution model, even if it seems to work well, cannot give correct  $\delta q$  values. The main reason is that these values arise from an extrapolation to infinite concentration. It is also concluded that  $\delta p$  values are reliable and comparison between IR and dielectric measurements show that  $\delta q$  is close to zero for LiSCN. Consequently, quadrupoles must be nonpolar in agreement with the result of the present paper.

D. Saar, J. Brauner, H. Farber, S. Petrucci

#### REFERENCES

1. H. Farber and S. Petrucci, J. Phys. Chem. 79, 1221 (1975).
2. H. Farber and S. Petrucci, J. Phys. Chem. 80, 327 (1976).
3. D. Saar, J. Brauner, H. Farber, S. Petrucci, J. Phys. Chem. 82, 545 (1978).
4. D. Menard and M. Chabanel, J. Phys. Chem. 79, 1081 (1975).
5. S. K. Garg and C. P. Smyth, J. Chem. Phys. 42, 1397 (1965).
6. S. Petrucci, P. Hemmes and M. Battistini, J. Am. Chem. Soc. 89, 5552 (1967).
7. J. P. Badiali, H. Cachet, A. Cyrot and J. C. Lestrade, J. Chem. Soc. Farad. Trans. II, 69, 1389 (1973).
8. C. F. J. Böttcher, "Theory of Electrical Polarization," Elsevier, Amsterdam (1973).
9. P. Jagodzinski and S. Petrucci, J. Phys. Chem. 78, 917 (1974); H. Farber and S. Petrucci, J. Phys. Chem. 80, 327 (1976); H. C. Wang and P. Hemmes, J. Am. Chem. Soc. 95, 5119 (1973).
10. R. M. Fuoss and M. Accascina, "Electrolytic Conductance," Intersci. Chapter XVIII, page 254 (1959).
11. M. Davies in "Dielectric Properties and Molecular Behavior," Edited by N. Hill, et al., Nostrand Reinhold Co. Ltd., London, page 298 (1969).
12. P. Debye, "Polar Molecules," Chemical Cat. Co., N.Y. (1929).
13. M. Chabanel, C. Menard, G. Guiheneuf, Compt. Rend. Acad. Sci., Paris, 272, 253 (1971).
14. D. Paoli, M. Lucon and M. Chabanel, to be published.

## CLASSICAL WAVE-PARTICLE DUALITY

N. Marcuvitz

A. Introduction

A detailed or microscopic analysis of wave propagation in dispersive nonlinear and/or turbulent systems is possible only in special cases and even then with difficulty. However, if waves are localizable in space-time as either coherent quasi-periodic wavepackets or as weakly correlated stochastic waves, then many average or macroscopic features of such wave systems can be analyzed by relatively simple analytical procedures.<sup>1</sup> These "averaging" procedures for the analysis of wave dynamics often admit simple physical interpretations in terms of the dynamics of "particles" of different types. Such interpretations are indicative of a wave-"particle" duality that is well known in quantum dynamics and is not completely unfamiliar in classical wave dynamics where, at least at linear level, it is often expressed in "ray-optic" language.<sup>2,3,4</sup> In quantum mechanics one views real particle motion statistically in terms of a probabilistic wave dynamics that on the average reproduces the classical dynamics of the particles. Conversely, in appropriate space-time regions one can view the average position, wavenumber spread, etc., of classical wavepackets in terms of the dynamics of corresponding "quasi-particle" systems. "Particle" systems may be described in terms of variables that are either kinetic (velocity, space and time dependent), fluid-dynamic (space and time dependent), or macroparticle (time-dependent) with increasing roughness of description. These different descriptions are related, of course, and are derivable from the kinetic by taking either velocity, or velocity and spatial moments of the kinetic distribution function. The macroparticle description provides a particularly simple means for illustrating particle-wave duality and will be employed both to review a quantum statistical interpretation of real particle motion in terms of probabilistic waves, and conversely to describe a localized classical wave field in terms of the dynamics of a macro-quasiparticle.

B. Particle-Wave Duality

At quantum levels vacuum fluctuation effects must be considered and hence the classical motion of a real particle may be viewed as a statistical average of a microscopic dynamics. More specifically, if a real particle of unit mass, with position  $q$  and momentum  $p$  at time  $t$  is described by a time dependent energy hamiltonian

$$H(p, q) = \frac{p^2}{2} + V(q) \quad (1)$$

then its classical motion, defined by the trajectory equations,

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q}, \quad \frac{dH}{dt} = 0, \quad (2)$$

is to be viewed in a statistical sense. Thus, one defines a probability density  $|\psi(q, t)|^2$  of finding the particle at position  $q$  at time  $t$ , with the complex wave function satisfying the probability normalization requirement

$$1 = \int_{-\infty}^{+\infty} \psi^*(q, t) \psi(q, t) dq, \quad \text{or} \quad 1 = (\psi, \psi) \quad (3)$$

and with the average particle position at time  $t$  given by

$$\bar{q}(t) = \int_{-\infty}^{+\infty} \psi^*(q, t) q \psi(q, t) dq, \quad \text{or} \quad \bar{q} = (\psi, Q\psi) \quad (4)$$

The right hand expressions are conventional hermitean inner product representations of those on the left;  $Q$  is a time independent "position operator" with representative  $q$  in the  $\psi$  basis. Similarly, one defines a probability density  $|\phi(p, t)|^2$  for the particle to have momentum  $p$  at time  $t$ , with the complex wave function  $\phi(p, t)$  also satisfying a probability normalization requirement

$$1 = \int_{-\infty}^{+\infty} \phi^*(p, t) \phi(p, t) dp, \quad \text{or} \quad 1 = \{\phi, \phi\}, \quad (5)$$

and hence with the average particle momentum at time  $t$  given by

$$\bar{p}(t) = \int_{-\infty}^{+\infty} \phi^*(p, t) p \phi(p, t) dp, \quad \text{or} \quad \bar{p} = \{\phi, P, \phi\}, \quad (6)$$

with  $P$  a "momentum operator" whose representative in the  $\phi$  basis is  $p$ . The bracketed expressions on the right of Eqs. (5) and (6) are evidently hermitean inner product representations of those on the left, but in  $p$  space rather than the  $q$  space products in Equations (3) and (4).

From Eqs. (3) and (5), using the Fischer-Riesz theorem,<sup>5</sup> one infers the existence in  $q$  space of an orthonormal basis  $\psi_p(q, t)$  with respect to which the wave function  $\phi(p, t)$  is the transform of  $\psi(q, t)$ , viz:

$$\psi(q, t) = \int_{-\infty}^{+\infty} \phi(p, t) \psi_p(q, t) dp \quad (6a)$$

with

$$\delta(p-p') = \int_{-\infty}^{+\infty} \psi_p^*(q, t) \psi_{p'}(q, t) dq, \quad \text{or} \quad \delta(p-p') = (\psi_p, \psi_{p'}) \quad (6b)$$

representative of the orthonormality properties of the  $\psi_p(q, t)$ . One observes that Eqs. (6a) and (6b) imply the normalization requirements

$$(\psi, \psi) = \{\phi, \phi\} = 1 \quad . \quad (6c)$$

Since  $\phi(p, t)$  is the transform of  $\psi(q, t)$ , then  $p\phi(p, t)$  must be the transform of some  $q$ -space wavefunction that can be identified as:

$$P\psi(q, t) = \int_{-\infty}^{+\infty} p\phi(p, t)\psi_p(q, t)dp \quad . \quad (7)$$

This identification in terms of the momentum operator  $P$  follows from the relation, deducible from Eqs. (6) and (7),

$$(\psi, P\psi) = \{\phi, P\phi\} = \bar{p}(t) \quad , \quad (8)$$

where the as yet unknown representative of the operator  $P$  in the  $\psi$  basis is to be distinguished from its known representative  $p$  in the  $\phi$  basis. Incidentally, one notes from Eq. (7) that the  $\psi_p$  are  $q$ -space eigenfunctions of the operator  $P$  with eigenvalues  $p$ .

Equations (4) and (8) not only identify the classical position  $\bar{q}(t)$  and momentum  $\bar{p}(t)$  variables as probabilistic averages of the  $\psi$  wave field, but also lend significance to the usual quantum mechanic postulational association of time independent operators and classical variables. A further observation, following from the reality of the classical variables is that these operators are hermitean, i.e.,

$$\begin{aligned} \bar{q}(t) &= (\psi, Q\psi) = (Q\psi, \psi) \\ \bar{p}(t) &= (\psi, P\psi) = (P\psi, \psi) \quad . \end{aligned} \quad (9a)$$

More generally, one associates the energy operator  $H(P, Q)$  with the average classical hamiltonian, via

$$\overline{H(p, q)} = (\psi, H(P, Q)\psi) \equiv (\psi, H\psi) = (H\psi, \psi) \quad . \quad (9b)$$

A defining equation for  $\psi$  may now be determined by imposition of the requirement that the classical equations (2) are "average" equations, whence from Eq. (9b)

$$\begin{aligned} \frac{d}{dt} \overline{H(p, q)} &= 0 = \left( \frac{\partial \psi}{\partial t}, H\psi \right) + \left( \psi, H \frac{\partial \psi}{\partial t} \right) \\ 0 &= \left( \frac{\partial \psi}{\partial t}, H\psi \right) + \left( H\psi, \frac{\partial \psi}{\partial t} \right) \quad . \end{aligned} \quad (10)$$

Equation (10) then implies that ( $i$  = imaginary unit)

$$H\psi \sim i \frac{\partial \psi}{\partial t}$$

is a possible (but not unique) defining equation for  $\psi$ . In quantum mechanics the proportionality constant is the Planck constant  $\hbar$ , and hence

$$H(P, Q)\psi = \hbar i \frac{\partial \psi}{\partial t} \quad , \quad (11)$$

becomes the Schrödinger equation defining the wave function  $\psi(q, t)$ . However, since the operator  $P$  is as yet undefined in the  $\psi$  basis, we observe from the interpretation of Eqs. (2) and (1) as averages that

$$\frac{d}{dt} \bar{q}(t) = \frac{\partial \overline{H(p, q)}}{\partial p} = \bar{p}(t)$$

whence from Eqs. (9a) and (11)

$$\left( \frac{\partial \psi}{\partial t}, Q\psi \right) + \left( \psi, Q \frac{\partial \psi}{\partial t} \right) = (\psi, P\psi)$$

or

$$\left( \psi, \frac{HQ - QH}{\hbar i} \psi \right) = (\psi, P\psi) \quad .$$

Using the operator expression for the hamiltonian, following from Eq. (1), one has

$$\frac{i}{2\hbar} (P^2 Q - Q P^2) = P \quad ,$$

which is satisfied if

$$PQ - QP = \frac{\hbar}{i} \quad ,$$

and hence in the  $\psi$  basis one infers the desired  $P$  representation:

$$P \rightarrow \frac{\hbar}{i} \frac{\partial}{\partial q} \quad , \quad \text{and} \quad Q \rightarrow q \quad . \quad (12)$$

Thus the Schrödinger equation (11) takes the more explicit form

$$H\left(\frac{\hbar}{i} \frac{\partial}{\partial q}, q\right)\psi = \frac{\hbar}{i} \frac{\partial \psi}{\partial t} \quad , \quad (13a)$$

or for  $H$  defined as in Eq. (1),

$$\left[ -\frac{\hbar^2}{2} \frac{\partial^2}{\partial q^2} + V(q) \right] \psi = \frac{\hbar}{i} \frac{\partial \psi}{\partial t} \quad (13b)$$

defines the wavefunction  $\psi(q, t)$  explicitly.

From the above derivation of Eq. (13) for  $\psi$ , it is manifest that the wave field  $\psi$  has been defined so that the averages Eq. (9a) reproduce the real classical motion defined by Eqs. (2). Since the  $\psi$  field contains more information than the classical equations, one can deduce from  $\psi$  other (quantum) properties of the particle motion such as indeterminacy relations for the mean square fluctuations of particle position and momentum, etc. Also, for the subsequent discussion, it is of interest to note that if one introduces in the  $\psi$  basis time independent operators  $\dot{Q}$ ,  $\dot{P}$  via:

$$\frac{d\bar{q}(t)}{dt} = (\psi, \dot{Q} \psi) , \quad \frac{d\bar{p}(t)}{dt} = (\psi, \dot{P} \psi) ,$$

then from Eqs. (11), et seq., one infers dynamical (operator) equations of motion

$$\begin{aligned} \dot{Q} &= \frac{i}{\hbar} [H, Q] \\ \dot{P} &= \frac{i}{\hbar} [H, P] , \end{aligned} \tag{14}$$

where  $[H, Q]$  is the commutator  $HQ - QH$ , etc. Equations (14) are of interest not only in the time dependent basis  $\psi$  (Schrödinger) but also in the time independent Heisenberg basis wherein  $P$  and  $Q$  are time dependent operators.

### C. Wave-Particle Duality

The average properties of a localized single-mode classical wavefield can be described in terms of the dynamics of a "macro quasi-particle" in a manner converse to that in Section B. By a macro quasi-particle is meant a localized space-time wave structure whose dynamics is characterized by position, "momentum," and "energy" variables similar to, but distinct from, those descriptive of a real particle. From a mathematical viewpoint if the wavefield is given, the macro quasi-particle description is unique, in contrast to the non uniqueness of probabilistic wave descriptions of a given real particle dynamics.

In appropriate ranges an overall wavefield may be decomposed into local mode equations of say the form

$$i \frac{\partial \psi}{\partial t} = H \left( \frac{1}{i} \frac{\partial}{\partial x} , x, |\psi|^{2n} \right) \psi = H\psi \tag{15}$$

where  $\psi = \psi(x, t)$  is a complex wave function and  $H$  is in general an  $x, t$  dependent non-linear hermitean operator (in  $x$ -space) indicative of the "energy" or frequency of the mode in question. Such decomposition into uncoupled equations of the type Eq. (15) is not always possible, in which case a multi-mode wavepacket description is required.

For nonlinear wavefields the decomposition procedure is, in general, nontrivial but its consideration will not be pursued further at this point.

For  $H$  an hermitean operator in  $x$ -space, local mode equations of the type Eq. (15) imply that the wave function  $\psi$  possesses the property

$$\frac{\partial}{\partial t} (\psi, \psi) = 0 \quad (16)$$

where, as in Section B, the hermitean inner product  $(\psi, \psi)$  is normalized to unity, with  $\psi \rightarrow 0$  as  $|x| \rightarrow \infty$ . A time dependent average property  $\bar{u}(t)$  of the wave function  $\psi$  is obtained from a classical hermitean operator  $U$  via:

$$\bar{u}(t) = (\psi, U\psi) \quad (17)$$

If the operator  $U$  is both  $x$  and  $t$  dependent, the time derivation of  $\bar{u}$  is given by

$$\frac{d}{dt} \bar{u}(t) = \left( \psi, \frac{\partial \psi}{\partial t}, U\psi \right) + \left( \psi, U \frac{\partial \psi}{\partial t} \right) + \left( \psi, \frac{\partial U}{\partial t} \psi \right)$$

whence in view of Equations (1) and (3)

$$\begin{aligned} \frac{d}{dt} \bar{u}(t) &= \left( \psi, \frac{HU - UH}{i} \psi \right) + \left( \psi, \frac{\partial U}{\partial t} \psi \right) \\ &= i[H, U] + \frac{\partial \bar{u}}{\partial t} \end{aligned} \quad (18)$$

For prescribed  $H$ , Eq. (18) leads to "equations of motion" for the average properties of a localized wavepacket whose exact space-time evolution is described by Equation (15).

For example, let  $\psi(x, t)$  be described by the linear wave equation

$$i \frac{\partial \psi}{\partial t} = \left( -\frac{1}{2} \frac{\partial^2}{\partial x^2} + V(x) \right) \psi = H\psi \quad (19)$$

with the initial condition  $\psi = \psi(x, 0)$ . Let us define  $k = \frac{1}{i} \frac{\partial}{\partial x}$  as a wave momentum operator and

$$\bar{x}(t) = (\psi, x\psi) \text{ and } \bar{k}(t) = \left( \psi, \frac{1}{i} \frac{\partial}{\partial x} \psi \right) \quad (20)$$

as the average position and average "momentum" of a macro quasi-particle descriptive of the wavepacket  $\psi(x, t)$ . The identification of  $\bar{k}(t)$  as the average momentum of the wavepacket follows from writing  $\psi = a(x, t) \exp[i\phi(x, t)]$ , whence  $\bar{k}(t) = (a, \frac{\partial \phi}{\partial x} a)$  since  $a \rightarrow 0$  as  $|x| \rightarrow \infty$ . The trajectory equations and other dynamical properties of the macro-"particle" can be inferred from the following commutator relations between

the  $k$  operator and an arbitrary function  $w = w(x)$  of the operator  $x$ , viz:

$$i[k, w] = w'(x) \quad (21a)$$

$$i[k^2, w] = kw'(x) + w'(x)k \quad (21b)$$

$$i[k^2, kw + wk] = k^2 w'(x) + w'(x)k^2 + 2kw'(x)k \quad (21c)$$

$$i[kx + xk, w] = 2xw'(x) \quad (21d)$$

$w'(x)$  is the  $x$  derivative of  $w(x)$ . Thus from Eqs. (21b) and (21a) with  $w = x$  and  $w = V(x)$ , respectively, one obtains from Eq. (18), with  $H = \frac{k^2}{2} + V(x)$ , the trajectory equations:

$$\frac{d\bar{x}}{dt} = \bar{k} \quad \frac{d\bar{k}}{dt} = -\frac{dV(\bar{x})}{d\bar{x}} \quad (22)$$

subject to initial conditions  $\bar{x}(0)$ ,  $\bar{k}(0)$  derivable from  $\psi(x, 0)$  and Equation (20). Similarly, for the higher moment averages:

$$\overline{x(t)^2} = (\psi, x^2 \psi), \quad \overline{k(t)^2} = (\psi, k^2 \psi), \quad \overline{kw + wk} = (\psi, (kw + wk) \psi),$$

one infers from Eqs. (18) and (21) the defining equations

$$\frac{d\bar{x}^2}{dt} = kx + xk \quad \frac{d\bar{k}^2}{dt} = kV'(x) + V'(x)k \quad (23)$$

$$\frac{d}{dt} (\overline{kx + xk}) = 2\bar{k}^2 - 2xV'(x)$$

For potential functions  $V(x)$  of higher than the second degree, Eqs. (22) and (23) are not closed in that the first order moments  $\bar{x}(t)$ ,  $\bar{k}(t)$  are dependent on higher moments  $\overline{x(t)^2}$ ,  $\overline{k(t)^2}$ , etc.

As a simple example, let  $V(x) = \frac{1}{2} \alpha^2 x^2$ , whence Eqs. (22) become

$$\frac{d\bar{x}}{dt} = \bar{k} \quad \frac{d\bar{k}}{dt} = -\alpha^2 \bar{x} \quad (24a)$$

whose solution is

$$\bar{x}(t) = \bar{x}(0) \cos \alpha t + \frac{\bar{k}(0)}{\alpha} \sin \alpha t \quad (24b)$$

$$\bar{k}(t) = \bar{k}(0) \cos \alpha t - \alpha \bar{x}(0) \sin \alpha t$$

Correspondingly, Eqs. (23) take the form:

$$\frac{d\overline{x^2}}{dt} = \overline{kx + xk} \quad \frac{d\overline{k^2}}{dt} = -\alpha^2 \frac{d\overline{x^2}}{dt} \quad (25a)$$

$$\frac{d}{dt} (\overline{kx + xk}) = \overline{2k^2} - 2\alpha^2 \overline{x^2}$$

which imply that  $\overline{k^2} + \alpha^2 \overline{x^2}$  is a constant of the motion and lead, for instance, to

$$\overline{x(t)^2} = \frac{1}{2} \left( \overline{x^2} - \frac{\overline{k^2}}{\alpha^2} \right)_0 \cos 2\alpha t + \left( \overline{kx + xk} \right)_0 \frac{\sin 2\alpha t}{2\alpha} + \frac{1}{2} \left( \overline{x^2} + \frac{\overline{k^2}}{\alpha^2} \right)_0 \quad (25b)$$

$$\overline{kx + xk} = \left( \overline{kx + xk} \right)_0 \cos 2\alpha t - \left( \overline{x^2} - \overline{k^2}/\alpha^2 \right)_0 \alpha \sin 2\alpha t$$

where the zero subscript denotes the value at  $t=0$ . The time-dependent "variances"  $\tilde{x}^2 = \overline{x^2} - \overline{x}^2$ ,  $\tilde{k}^2 = \overline{k^2} - \overline{k}^2$ , etc. are useful measures of the evolving shape of the macroparticle and can be readily evaluated from Equations (24) and (25)

A non-closed example is provided by the nonlinear space-time dependent potential  $V(x, t) = \beta |\psi|^2$  descriptive of the nonlinear Schrödinger equation, which is characterized by a time dependent hermitean operator  $H = k^2/2 + \beta |\psi|^2$ . Equations (22) still apply and become

$$\frac{d\overline{x}}{dt} = \overline{k} \quad \frac{d\overline{k}}{dt} = -\beta \int |\psi|^2 \frac{\partial}{\partial x} |\psi|^2 dx = 0, \quad (26)$$

whence

$$\overline{x}(t) = \overline{x}(0) + \overline{k}(0)t \quad \overline{k}(t) = \overline{k}(0)$$

Equation (26) implies that a wavepacket having an initially symmetric amplitude and phase with  $\overline{x}(0) = \overline{k}(0) = 0$  will continue to maintain this symmetry as time evolves. The equations (23) for the higher moments lead to a nonclosed description of the macroparticle dynamics in that they involve such moments as (let  $\psi = ae^{i\phi}$ )

$$\int |\psi|^4 dx = \int a^4 dx$$

and

$$\frac{1}{i} \int \psi^* \frac{\partial \psi}{\partial x} - \psi \frac{\partial \psi^*}{\partial x} \frac{\partial |\psi|^2}{\partial x} dx = 2 \int \frac{\partial \phi}{\partial x} \frac{\partial a^2}{\partial x} dx$$

whose evolution cannot be readily ascertained unless one solves the defining equation for  $\psi$  in detail. Specifically, Eqs. (23) take the form

$$\begin{aligned}\overline{\frac{dx}{dt}}^2 &= \overline{kx + xk} & \frac{d}{dt} (\overline{kx + xk}) &= \overline{2k^2} - \beta \int a^4 dx \\ \overline{\frac{dk}{dt}}^2 &= -\beta \int \frac{\partial \phi}{\partial x} \frac{\partial a^4}{\partial x} dx\end{aligned}\quad (27)$$

which are obviously not closed. From the defining Eq. (19) for  $\psi$  one infers that

$$\frac{d}{dt} \int a^4 dx = \int \frac{\partial \phi}{\partial x} \frac{\partial a^4}{\partial x} dx \quad (28a)$$

$$\frac{d}{dt} \int \frac{\partial \phi}{\partial x} \frac{\partial a^4}{\partial x} dx = -2 \int \left[ \left( a^2 \frac{\partial^2 \phi}{\partial x^2} \right)^2 + \beta \left( a \frac{\partial a^2}{\partial x} \right)^2 - \frac{1}{4} \frac{\partial^2 a^4}{\partial x^2} \frac{1}{a} \frac{\partial^2 a}{\partial x^2} \right] dx \quad (28b)$$

Equation (28a) permits one to infer from Eq. (27) that

$$\begin{aligned}\overline{\frac{d^2 x}{dt^2}}^2 &= \overline{2k^2} - \beta \int a^4 dx \\ \frac{d}{dt} (\overline{k^2} + \beta \int a^4 dx) &= 0\end{aligned}\quad (29)$$

Equations (26) and (29) yield interesting but limited information about the evolution of the initial wavepacket. Further information is dependent either on the determination of the time dependence of  $\int a^4 dx$ , a non-closed problem as noted in Eqs. (28), or on equivalent kinetic information discussed elsewhere.

Office of Naval Research  
N00014-76-C-0176

N. Marcuvitz

#### REFERENCES

1. G. B. Whitham, "Linear and Nonlinear Waves," Wiley Interscience (1974).
2. S. N. Vlasov, V. A. Petrishev and V. I. Talanov, "Averaged Description of Wave Beams," *Radiofizika*, Vol. 14, No. 9, pp. 1353-63 (September 1971).
3. M. C. Newstein and D. Ramakrishnan, "Application of Heisenberg Picture to Averaged Description of the Propagation of Optical Beams," Report No. POLY-MRI-1394-78 (June 1978).
4. B. B. Kadomtsev and V. I. Karpman, "Nonlinear Waves," *Soviet Physics Uspekhi*, Vol. 14, No. 1 (July 1973).
5. R. Courant and D. Hilbert, "Methods of Mathematical Physics," Interscience, Vol. 1, Chapter 2, p. 110 (1953).

## WAVEPACKETS AS QUASIPARTICLE SYSTEMS

N. Marcuvitz

In nonlinear and/or turbulent wave propagation it is sometimes useful to view a wavepacket as a system of point quasiparticles, each described by its position  $\mathbf{x}(t)$ , momentum  $\mathbf{k}(t)$ , and energy  $\omega(\mathbf{k}, \mathbf{x}, t)$ . The individual quasiparticle motion, which is determined by the functional form of  $\omega$  and differs with each mode type, provides a physical picture of how a wavepacket evolves in space-time. A kinetic description of the overall quasiparticle system at time  $t$  is given in terms of a distribution function  $F(\mathbf{k}, \mathbf{x}, t)$  in  $\mathbf{k}, \mathbf{x}$  space. A fluid description is obtained from the  $\mathbf{k}$  moments of  $F(\mathbf{k}, \mathbf{x}, t)$  and a macro-quasiparticle description is recovered from the  $\mathbf{k}$  and  $\mathbf{x}$  moments of  $F(\mathbf{k}, \mathbf{x}, t)$ .

Consider a single mode type of wave process described locally by a complex wave function  $\psi(\mathbf{x}, t)$  satisfying

$$i \frac{\partial \psi}{\partial t} = H \psi \quad (1)$$

with

$$H = H\left(\frac{1}{i} \frac{\partial}{\partial \mathbf{x}}, \mathbf{x}, t\right)$$

subject to an initial condition  $\psi(\mathbf{x}, 0)$ . To view the time evolution of  $\psi$  in terms of a system of quasiparticles, it is necessary to identify the quasiparticle descriptors as well as the kinetic distribution function  $F$  from the above defining equation for  $\psi$ . For this purpose one introduces the double Fourier representation.

$$\psi_1 \psi_2^* = \iint \phi_1 \phi_2^* \exp[i(k_1 x_1 - k_2 x_2)] dk_1 dk_2 / (2\pi)^2 \quad (2)$$

where  $\phi_1 = \phi(k_1, t)$  and  $\phi_2 = \phi(k_2, t)$  are, respectively, the Fourier amplitudes of  $\psi_1 = \psi(x_1, t)$  and  $\psi_2 = \psi(x_2, t)$ . The product  $\psi_1 \psi_2^*$  displays both a fast and slow spatial behavior and it is desirable to average over the very fast behavior. If the  $\psi(\mathbf{x}, t)$  and hence the  $\phi(\mathbf{k}, t)$  are stochastic, the averaging is explicit with Eq. (2) being expressed in terms of ensemble averages as

$$\langle \psi_1 \psi_2^* \rangle = \iint \langle \phi_1 \phi_2^* \rangle \exp[i(k_1 x_1 - k_2 x_2)] dk_1 dk_2 / (2\pi)^2 \quad (2a)$$

and the average symbol  $\langle \rangle$  should be retained throughout the following. However, if the  $\psi$  and  $\phi$  are deterministic, this symbol should be understood as an average over a suitably fast spatial period or phase and will be omitted. After the averaging, "slow" ( $\mathbf{x}$  or  $\mathbf{K}$ ) and "fast" ( $\xi$  or  $\mathbf{k}$ ) spatial or wavenumber variables are defined by:

$$x = (x_1 + x_2)/2 \quad K = k_1 - k_2 \quad (3)$$

and

$$\xi = x_1 - x_2 \quad k = (k_1 + k_2)/2$$

whence since

$$k_1 x_1 - k_2 x_2 = k\xi + Kx \quad \text{and} \quad dk_1 dk_2 = dk dK,$$

Eq. (2) can be rewritten as

$$\psi_1 \psi_2^* = \int F(k, x, t) \exp(ik\xi) dk / 2\pi \quad (3a)$$

where

$$F(k, x, t) = \int \phi_1 \phi_2^* \exp(iKx) dK / 2\pi \quad (3b)$$

The reality property  $F(k, x, t) = F^*(k, x, t)$  of the fast (correlation) spectrum of  $\psi_1 \psi_2$  follows readily from Eq. (3); however, positive definiteness is not in general assured. Fourier transformations inverse to Eq. (3) are manifestly

$$F(k, x, t) = \int \psi_1 \psi_2^* \exp(ik\xi) d\xi \quad (4a)$$

$$\phi_1 \phi_2^* = \int F(k, x, t) \exp(-iKx) dx \quad (4b)$$

For  $\xi = 0$  and  $K = 0$ , Eqs. (3a) and (4b) yield

$$|\psi(x, t)|^2 = \int F(k, x, t) dk / 2\pi$$

$$|\phi(k, t)|^2 = \int F(k, x, t) dx$$

whence one infers that

$$\int |\psi(x, t)|^2 dx = \int |\phi(k, t)|^2 dk / 2\pi = \int F(k, x, t) dk dx / 2\pi \quad (5)$$

Even though only reality but not positive definiteness of  $F(k, x, t)$  is assured, Eq. (5) suggests the identification of  $F(k, x, t)$  as a number density of quasiparticles in  $k, x$  phase space. The consequent interpretation of  $|\psi|^2$  as a quasiparticle number density in real space and of  $|\phi|^2$  as a number density in momentum space are in conformity with a similar concept in quantum mechanics although in the latter instance because of a different normalization  $|\psi|^2$  and  $|\phi|^2$  are given a probabilistic interpretation.

The defining equation for  $F(k, x, t)$  can be deduced from Eq. (1) if one infers therefrom

$$i \frac{\partial}{\partial t} (\psi_1 \psi_2^*) = (H_1 - H_2^*) (\psi_1 \psi_2^*) \quad (6)$$

where

$$H_1 = H\left(\frac{1}{i} \frac{\partial}{\partial x_1}, x_1, t\right) \quad H_2^* = H\left(-\frac{1}{i} \frac{\partial}{\partial x_2}, x_2, t\right)$$

Noting the definitions Eq. (3) and forming the inverse transform Eq. (4a) of Eq. (6), one obtains

$$i \frac{\partial F}{\partial t} = \int \exp(-ik\xi) \left[ H\left(\frac{1}{i} \frac{\partial}{\partial \xi} + \frac{1}{2i} \frac{\partial}{\partial x}, x + \frac{\xi}{2}, t\right) - H\left(\frac{1}{i} \frac{\partial}{\partial \xi} - \frac{1}{2i} \frac{\partial}{\partial x}, x - \frac{\xi}{2}, t\right) \right] \psi_1 \psi_2^* d\xi.$$

After commuting the operator  $H$  so that it acts only on  $\exp(-ik\xi)$  with a consequent change in sign of  $\frac{\partial}{i\partial \xi}$ , one finds as the defining "kinetic" equation for  $F$

$$i \frac{\partial F}{\partial t} = \left[ H\left(k + \frac{1}{2i} \frac{\partial}{\partial x}, x - \frac{1}{2i} \frac{\partial}{\partial k}, t\right) - H\left(k - \frac{1}{2i} \frac{\partial}{\partial x}, x + \frac{1}{2i} \frac{\partial}{\partial k}, t\right) \right] F \quad (7)$$

Assuming a weak dependence of  $F$  on  $k$  and  $x$ , one expands  $H$  in first order to obtain as the "quasiparticle approximation" to (7):

$$\frac{\partial F}{\partial t} + \frac{\partial \omega}{\partial k} \frac{\partial F}{\partial x} - \frac{\partial \omega}{\partial x} \frac{\partial F}{\partial k} = 0 \quad (8)$$

where  $\omega = H(k, x, t)$ . The characteristic equations of the partial differential Eq. (8), namely:

$$\frac{dx}{dt} = \frac{\partial \omega}{\partial k}, \quad \frac{dk}{dt} = -\frac{\partial \omega}{\partial x} \quad (9)$$

identify the trajectories of quasiparticles of position  $x(t)$ , momentum  $k(t)$  and energy  $\omega(k, x, t)$ . On a quasiparticle trajectory, one observes in the quasiparticle approximation that for the mode type Eq. (1):

$$\frac{dF}{dt} = 0 \quad \text{and} \quad \frac{d\omega}{dt} = \frac{\partial \omega}{\partial t} \quad (10)$$

The implied constancy of  $F$  on a quasiparticle trajectory starting at  $x(0) = x_0$  and  $k(0) = k_0$  permits the solution of Eq. (8), evolving from a given initial state  $F(k, x, 0)$ , to be written conventionally as

$$F(k, x, t) = F(k_0(k, x, t), x_0(k, x, t), 0) \quad (11)$$

where  $k_0(k, x, t)$ ,  $x_0(k, x, t)$  define the quasiparticle trajectory starting at  $k_0$ ,  $x_0$ , at time 0 and ending at  $k, x$  at time  $t$ .

"Fluid dynamic" properties of the quasiparticle system can be inferred from the  $k$ -moments of the kinetic distribution function  $F$ . Such properties are representative of macroscopic characteristics of wavepacket solutions of Eq. (1) in regions wherein wavepackets form.

If a typical wavepacket solution is represented as

$$\psi(x, t) = a(x, t) \exp(i\theta(x, t)) \quad (12)$$

then as noted in Eq. (5) the local amplitude  $a$  follows from  $F(k, x, t)$  via:

$$a^2(x, t) = \int F(k, x, t) dk / 2\pi \quad (13)$$

One observes that the amplitude squared  $a^2(x, t)$  of the wavepacket, or the zeroth moment of  $F$ , is indicative of the number density of quasiparticles. The local phase  $\theta(x, t)$  of the wavepacket can be determined from the local wavenumber and frequency,

$$\bar{k}(x, t) = \frac{\partial \theta}{\partial x} \quad \text{and} \quad \bar{\nu}(x, t) = -\frac{\partial \theta}{\partial t}, \quad (14a)$$

by noting that

$$\theta(x, t) = \theta(x_0, 0) + \int (\bar{k}dx - \bar{\nu}dt) \quad (14b)$$

where the integration in  $x, t$  space is over an appropriate path extending from  $x_0, 0$  to  $x, t$ . To find the local wavenumber and frequency one first observes, in view of (14a), that

$$\bar{k} = (\psi^* \frac{\partial}{\partial x} \psi - \psi \frac{\partial}{\partial x} \psi^*) / 2ia^2$$

$$\bar{\nu} = i(\psi^* \frac{\partial}{\partial t} \psi - \psi \frac{\partial}{\partial t} \psi^*) / 2a^2$$

whence, using Eqs. (3a) and (1), one derives the relations

$$\frac{1}{2} (\psi_2^* \frac{\partial}{\partial x_1} \psi_1 - \psi_1 \frac{\partial}{\partial x_2} \psi_2^*)_{\xi=0} = \frac{1}{2} \int k F(k, x, t) dk / 2\pi = \bar{k}(x, t) a^2 \quad (15a)$$

$$\frac{1}{2} (\psi_2^* H_1 \psi_1 - \psi_1 H_2^* \psi_2^*)_{\xi=0} = \int [\omega(k, x, t) - \frac{1}{8} \frac{\partial^2 \omega}{\partial k^2} \frac{\partial^2}{\partial x^2} + \dots] F dk / 2\pi = \bar{\nu}(x, t) a^2 \quad (15b)$$

which identify the local wavenumber  $\bar{k}$  and frequency  $\bar{\omega}$  as averages or moments of the quasiparticle distribution function  $F$ . Thus, via Eqs. (3), (8), (13), and (14), the space time dependent local characteristics of a propagating wavepacket can be ascertained from a kinetic description starting from initial values  $F(k, x, 0)$  and  $\theta(x, 0)$ , which are derivable from  $\psi(x, 0)$ .

An alternative "fluid dynamic" picture of a propagating wavepacket can be inferred from moments of the quasiparticle kinetic Equation (8). This "fluid" description is closely related to the conservation equations obtained by Whitham from his averaged variational formalism for propagating wavepackets. One defines a local velocity  $\bar{v}$  and force  $\bar{f}$  for the quasiparticle system representing the wavepacket via the  $F$  moments:

$$a^2 \bar{v}(x, t) = \int \frac{\partial \omega}{\partial k}(k, x, t) F(k, x, t) dk / 2\pi \quad (16)$$

$$a^2 \bar{f}(x, t) = \int \frac{\partial^2 \omega}{\partial k^2} \frac{\partial \omega}{\partial x}(k, x, t) F(k, x, t) dk / 2\pi$$

Then from the quasiparticle kinetic Eq. (8), written in the form:

$$\frac{\partial F}{\partial t} + \frac{\partial}{\partial x} \left( \frac{\partial \omega}{\partial k} F \right) - \frac{\partial}{\partial k} \left( \frac{\partial \omega}{\partial x} F \right) = 0,$$

one obtains, by forming in a conventional manner zeroth and velocity moments, the "fluid" equations:

$$\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x} (a^2 \bar{v}) = 0 \quad (17a)$$

$$\frac{\partial}{\partial t} (a^2 \bar{v}) + \frac{\partial}{\partial x} (a^2 \bar{v}^2) - a^2 \bar{f} = 0 \quad (17b)$$

Using Eq. (17b), one can rewrite Eq. (17b) in the more conventional "fluid" form

$$\frac{\partial \bar{v}}{\partial t} + \bar{v} \frac{\partial \bar{v}}{\partial x} + \frac{1}{a} \frac{\partial \bar{p}}{\partial x} = \bar{f} \quad (17b)$$

where, as in (16), one defines

$$a^2 \bar{v}^2 = \int \left( \frac{\partial \omega}{\partial k} \right)^2 F dk / 2\pi \quad (18a)$$

and introduces a local quasiparticle "pressure"  $\bar{p}(x, t)$  via

$$a^2 \bar{v}^2 = a^2 \bar{v}^2 + \bar{p} \quad (18b)$$

Equations (17) constitute two equations for four unknowns and hence additional information on  $p$  and  $f$  is required for closure. Determination of the requisite information is perhaps best clarified by a particular example.

For the case where the quasiparticle "energy" is specified by the dispersion relation

$$\omega(k, x, t) = \frac{k^2}{2} + V(x, t) \quad , \quad (19a)$$

corresponding to an Eq. (1) with

$$H = -\frac{1}{2} \frac{\partial^2}{\partial x^2} + V(x, t) \quad (19b)$$

Eqs. (16) imply that

$$\bar{v} = \bar{k} \quad \text{and} \quad \bar{f} = -\frac{\partial V}{\partial x} \quad . \quad (20a)$$

To determine the local "pressure"  $\bar{p}$  one first observes that

$$-\left(\psi_2^* \frac{\partial^2}{\partial x_1^2} \psi_1 + \psi_1 \frac{\partial^2}{\partial x_2^2} \psi_2^*\right) = \int_{\xi=0} \left[ \left(k + \frac{1}{2i} \frac{\partial}{\partial x}\right)^2 - \left(k - \frac{1}{2i} \frac{\partial}{\partial x}\right)^2 \right] F dk / 2\pi$$

whence evaluating the right hand side, inserting Eq. (12) into the left side and equating, one finds

$$\bar{k}^2 - \frac{1}{a} \frac{\partial^2 a}{\partial x^2} = \overline{k^2} - \frac{1}{4a} \frac{\partial^2}{\partial x^2} a^2 \quad ; \quad (20b)$$

hence, since  $\bar{v} = \bar{k}$  and  $\bar{v}^2 = \overline{k^2}$ , from (18b)

$$\bar{p} = \frac{1}{4} \frac{\partial^2}{\partial x^2} a^2 - a \frac{\partial^2}{\partial x^2} a \quad (20c)$$

and therefore

$$\frac{1}{a} \frac{\partial \bar{p}}{\partial x} = -\frac{\partial}{\partial x} \left( \frac{1}{2a} \frac{\partial^2 a}{\partial x^2} \right) \quad (20d)$$

On use of the identities Eqs. (20), (17a) and (17b), become a closed set of "fluid" equations for the local quasiparticle density  $a^2$  and velocity  $\bar{k}$ , viz:

$$\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x} (a^2 \bar{k}) = 0 \quad (21a)$$

$$\frac{\partial \bar{k}}{\partial t} + \bar{k} \frac{\partial \bar{k}}{\partial x} - \frac{\partial}{\partial x} \left( \frac{1}{2a} \frac{\partial^2 a}{\partial x^2} \right) = - \frac{\partial V}{\partial x} \quad (21b)$$

In the context of the Whitham formalism for the case Eq. (19), it is of interest to note from Eqs. (19a), (15b), and (20b) that the local frequency  $\nu$  takes the form

$$\bar{\nu} = \frac{\bar{k}^2}{2} - \frac{1}{2a} \frac{\partial^2 a}{\partial x^2} + V \quad (22)$$

and identifies the averaged dispersion relation in the Whitham formalism. The relation

$$\frac{\partial \bar{k}}{\partial t} + \frac{\partial \bar{\nu}}{\partial x} = 0 ,$$

implied by Eq. (14a) and termed the "wave conservation equation" by Whitham then becomes with the aid of Eq. (22) the "fluid momentum" Equation (21b).

Numerical comparisons of the quasiparticle approximations with exact wavepacket solutions will appear in a subsequent report.

Office of Naval Research  
N00014-76-C-0176

N. Marcuvitz

#### REFERENCES

1. G. B. Whitham, "Linear and Nonlinear Waves," Wiley-Interscience, (1974).
2. B. B. Kadomtsev and V. I. Karpman, "Nonlinear Waves," Soviet Physics Uspekhi, Vol. 14, No. 1, (July 1971).
3. V. N. Tsytovich, "Nonlinear Effects in Plasma," Plenum Press (1970).
4. F. D. Tappert and I. M. Besieris, "Stochastic Wave Kinetic Equation and Application to Wavepacket Spreading," International Symp. on Electromagnetic Wave Theory, Tbilisi, USSR, (September 1971).
5. N. Marcuvitz, "On Quasiparticle Description of Many Particle Systems," IEEE Trans. on Electron Devices, Vol. ED-17, No. 3, pp. 252-7 (March 1970).

## RENORMALIZATION OF MAXWELL'S EQUATIONS FOR TURBULENT PLASMA

S. Barone and N. Marcuvitz

The statistical theory of a classically describable plasma poses a coupled field problem governed by Maxwell's equations\*

$$\begin{aligned} \partial_t \hat{\underline{B}} + \nabla \times \hat{\underline{E}} &= 0, & \nabla \cdot \hat{\underline{B}} &= 0, \\ \partial_t \hat{\underline{D}} - \nabla \times \hat{\underline{H}} &= -\hat{\underline{J}} - \hat{\underline{J}}_{\text{ext}}, & \nabla \cdot \hat{\underline{D}} &= \hat{\rho} + \hat{\rho}_{\text{ext}}, \end{aligned} \quad (1)$$

and the Klimontovich equations

$$\left[ \partial_t + \underline{v} \cdot \nabla + \frac{q_\alpha}{m_\alpha} (\hat{\underline{E}} + \underline{v} \times \hat{\underline{B}}) \cdot \nabla_{\underline{v}} \right] \hat{f}_\alpha = \hat{\eta}_\alpha, \quad (2)$$

one for each species of charged particle ( $\alpha = 1, 2, \dots$ ). For any particular member of the statistical ensemble the current ( $\hat{\underline{J}}$ ) and charge ( $\hat{\rho}$ ) densities that couple Maxwell's equations to the Klimontovich equations are given by velocity integrals of the particle distribution functions for that member of the ensemble

$$\begin{aligned} \hat{\underline{J}}(\underline{r}, t) &= \sum_\alpha \int (d\underline{v}) q_\alpha \underline{v} \hat{f}_\alpha(\underline{v}, \underline{r}, t), \\ \hat{\rho}(\underline{r}, t) &= \sum_\alpha \int (d\underline{v}) q_\alpha \hat{f}_\alpha(\underline{v}, \underline{r}, t). \end{aligned} \quad (3)$$

The microscopic content of the Klimontovich equations is that the stochastic fields  $\hat{\underline{E}}$ ,  $\hat{\underline{B}}$  determine the particle orbits according to the usual Lorentz force law. It is convenient to include external current and charge densities  $\hat{\underline{J}}_{\text{ext.}}$ ,  $\hat{\rho}_{\text{ext.}}$ , the space-time dependence of which is prescribed, i.e., unaffected by the behavior of either the field vectors or the particle distribution functions. The external sources are, for example, used to generate externally imposed fields which are assumed to be unaffected by the motion of the field-plasma system and, also, to radiate the initial field distribution (at say,  $t=0$ ) from a quiescent situation for  $t < 0$ . The external sources  $\hat{\eta}_\alpha$  serve the same purposes for the particle systems.

The Maxwell-Klimontovich equations characterize an attempt to relate the macroscopic behavior of a field-plasma system to the behavior of single particles. If (as an approximation) it is assumed that the particles in any particular member of the ensemble move as if only the ensemble averaged field acted on them, the solution is secular.

\*For any quantity,  $\hat{X}$  defined on the ensemble we use  $X \equiv \langle \hat{X} \rangle$  to denote its average over the ensemble and  $\tilde{X}$  to denote the difference between  $\hat{X}$  for any particular member of the ensemble and the average, i.e.,  $\tilde{X} \equiv \hat{X} - X$ .

Thus, the effect of field fluctuations  $\tilde{\mathbf{E}}$ ,  $\tilde{\mathbf{B}}$  on the particle orbits is crucial in the development of the general theory. The analogy between the present situation and quantum electrodynamics is apparent. It is now widely believed that the macroscopic physical content of the Maxwell-Klimontovich equations will be more clearly revealed if these equations are renormalized in a manner similar to that which has been successful in electrodynamics. The expectation is that the major effect of the field fluctuations on the particle orbits can be identified and handled in a general way leaving only genuinely small residual effects to be treated by perturbation theoretic procedures.

In this work it is shown how Maxwell's equations can be completely renormalized without approximation or expansion. It is anticipated that this formulation of the theory will be useful when parametric and other nonlinear effects occur within the spectrum of the fluctuating field. Secondly, we emphasize the extent to which renormalization procedures are arbitrary and the circumstances under which our procedure makes optimum use of the available information. Thirdly, we derive a renormalized equation for the two point field correlation function. This equation is the analog for the electromagnetic field of the Bethe-Salpeter equation. It leads to kinetic equations for interacting quasi-particles. Finally, we discuss a Klimontovich equation for interacting quasi-particles. This equation is appropriate for the discussion of the collective modes of oscillation of quasi particles and their description in terms of new types of quasi particles.

Following Weinstock<sup>1</sup> we write the fluctuating part of the distribution function in the form

$$\tilde{f} = \hat{G}_A(\tilde{\eta} + \tilde{V}f) \quad (4)$$

where  $\hat{G}_A$  is an appropriately defined Green's function and  $\tilde{V}$  is the fluctuating part of the operator in Equation (2). The kinetic equation for  $f$  becomes

$$[L_0 - \langle \tilde{V} \hat{G}_A \tilde{V} \rangle] f = \eta + \langle \tilde{V} \hat{G}_A \tilde{\eta} \rangle, \quad (5)$$

where  $L_0$  is the ensemble average of the operator in Equation (2). This equation relates the ensemble averaged distribution function,  $f$ , to the source  $\hat{\eta}$ . The relationship between  $f$  and  $\hat{\eta}$  depends on the fluctuating fields,  $\tilde{V}$ .

The renormalization of Maxwell's equations is straightforward once the fluctuating part of the current density,  $\hat{\mathbf{j}}$ , is decomposed according to Equation (4). For simplicity of presentation we neglect the magnetic field fluctuations in  $\tilde{V}$  and specialize to a single component plasma ( $q \rightarrow -e$ ). Maxwell's equations may be written

$$Y_M \cdot \tilde{E} = -\tilde{J}_{\text{ext.}} + \int (d\underline{v}) e \underline{v} \hat{f}(\underline{v}) \quad , \quad (6)$$

where  $Y_M$  is the field operator discussed in Ref. (2) and the velocity coordinate has been made explicit. For the ensemble averaged field we have

$$Y_M \cdot \underline{E} = -\underline{J}_{\text{ext.}} + \int (d\underline{v}) e \underline{v} f(\underline{v}) \quad , \quad (7)$$

which is to be solved self consistently with Eq. (5) for the average distribution function,  $f$ . These two equations do not form a closed system because the equation for  $f$  involves the field fluctuations,  $\tilde{E}$ .

Using Eq. (4) the fluctuating part of the field is seen to satisfy

$$Y_M \cdot \tilde{E} = -\tilde{J}_{\text{ext.}} + \int e \underline{v} \hat{G}_A (\tilde{\eta} + \tilde{V} f) \quad , \quad (8)$$

where we have shortened the notation so that velocity sums are indicated by a simple integral sign. Since  $\tilde{V} = (e/m) \tilde{E} \cdot \nabla_{\underline{v}}$  the above equation may be written

$$\left[ Y_M - \frac{e^2}{m} \int \underline{v} \hat{G}_A \nabla_{\underline{v}} f \right] \cdot \tilde{E} = -\tilde{J}_{\text{ext.}} + \int e \underline{v} \hat{G}_A \tilde{\eta} \quad , \quad (9)$$

or

$$(Y_M + P + \hat{P}_A) \cdot \tilde{E} = -(\tilde{J}_{\text{ext.}} + \tilde{J}'_{\text{ext.}}) \quad , \quad (10)$$

where

$$\tilde{J}'_{\text{ext.}} \equiv \int (-e) \underline{v} \hat{G}_A \tilde{\eta} \quad , \quad (11)$$

and

$$-\hat{P}_A \equiv \frac{e^2}{m} \int \underline{v} \hat{G}_A \nabla_{\underline{v}} f + P \quad , \quad (12)$$

where  $P$  is an arbitrary operator that does not vary over the ensemble. According to the present point of view, renormalization is accomplished by selecting  $P$  so that the  $\hat{P}_A$  term in the above equation has as small an effect as possible on the evolution of  $\tilde{E}$ . If, in fact,  $\hat{P}_A$  has a small effect

$$(-Y_m - P) \cdot \tilde{E} \approx (\tilde{J}_{\text{ext.}} + \tilde{J}'_{\text{ext.}}) \quad , \quad (13)$$

which shows that  $P$  should be selected in such a way that the Green's function  $(-Y_m - P)^{-1}$  relates the fluctuating part of the field,  $\tilde{E}$  to the fluctuating part of the externally

externally imposed current densities  $\tilde{\mathbf{J}}_{\text{ext.}}$  and  $\tilde{\mathbf{J}}'_{\text{ext.}}$ . In other words  $(-Y_M - P)^{-1}$  should relate the deviation from the average field in any particular member of the ensemble to the deviation from the average of the externally imposed current densities. There is no one choice for  $P$  that will accomplish this exactly for all members of the ensemble. However, the relationship between  $\tilde{\mathbf{E}}$  and  $(\tilde{\mathbf{J}}_{\text{ext.}} + \tilde{\mathbf{J}}'_{\text{ext.}})$  will not vary by a large amount over most of the ensemble. The overwhelming majority of the members of the ensemble evolve in nearly the same way as the ensemble average. In the absence of further information the optimum choice for  $P$  is such that  $(-Y_M - P)^{-1}$  is the same as the Green's function that relates a small change in the external current density  $\delta \mathbf{J}_{\text{ext.}}$  to the corresponding small change in the ensemble average  $\delta \tilde{\mathbf{E}}$ . In particular, suppose that the small quantities  $\delta \tilde{\mathbf{E}}$  and  $\delta \mathbf{J}_{\text{ext.}}$  are linearly related by:

$$\delta \tilde{\mathbf{E}} = \mathcal{G}' \cdot \delta \mathbf{J}_{\text{ext.}}, \quad (14)$$

where  $\mathcal{G}'$  is an appropriate Green's function. We propose that  $P$  be selected so that

$$\frac{1}{-Y_M - P} = \mathcal{G}' \quad (15)$$

and in analogy to the corresponding situation in electrodynamics refer to  $P$  as the polarization operator.<sup>2</sup> In functional derivative notation

$$\mathcal{G}' = \frac{\delta \tilde{\mathbf{E}}}{\delta \mathbf{J}_{\text{ext.}}}, \quad (16)$$

or

$$P = \frac{\delta \mathbf{J}}{\delta \mathbf{J}_{\text{ext.}}} \cdot \mathcal{G}'^{-1} = \frac{\delta \mathbf{J}}{\delta \tilde{\mathbf{E}}}. \quad (17)$$

A renormalized differential equation characterization of the two-point field correlation function follows from

$$(Y_M + P) \cdot \tilde{\mathbf{E}} = -(\tilde{\mathbf{J}}_t + \tilde{\mathbf{J}}_1), \quad (18)$$

where  $\tilde{\mathbf{J}}_1 \equiv \hat{P}_A \tilde{\mathbf{E}}$ . In an obvious notation

$$(Y_M + P)_1 (Y_M + P)_2 : \langle \tilde{\mathbf{E}}(1) \tilde{\mathbf{E}}(2) \rangle = \langle [\tilde{\mathbf{J}}_t(1) + \tilde{\mathbf{J}}_1(1)] [\tilde{\mathbf{J}}_t(2) + \tilde{\mathbf{J}}_1(2)] \rangle, \quad (19)$$

so that defining an interaction operator,<sup>2</sup>  $I$  by

$$\langle \tilde{\mathbf{J}}_1(1) \tilde{\mathbf{J}}_1(2) \rangle \equiv I(1, 1'; 2, 2') : \langle \tilde{\mathbf{E}}(1') \tilde{\mathbf{E}}(2') \rangle, \quad (20)$$

we have

$$\begin{aligned} & \left[ (Y_M + P)_1 (Y_M + P)_2 - I_{12} \right] : \langle \tilde{E}(1) \tilde{E}(2) \rangle \\ &= \langle \tilde{J}_t(1) \tilde{J}_t(2) \rangle + \langle \tilde{J}_t(1) \tilde{J}_1(2) + \tilde{J}_1(1) \tilde{J}_t(2) \rangle \end{aligned} \quad (21)$$

If  $\tilde{J}_1$  is related back to the external currents  $\tilde{J}_t$  via an operator  $B_{12}$ , the above equation for the two-point field correlation function reads

$$\left[ (Y_M + P)_1 (Y_M + P)_2 - I_{12} \right] : \langle \tilde{E}(1) \tilde{E}(2) \rangle = B_{12} : \langle \tilde{J}_t(1) \tilde{J}_t(2) \rangle \quad (22)$$

If a two-particle Green's function is defined by

$$\left[ (Y_M + P)_1 (Y_M + P)_2 - I_{12} \right] : G_{12} = B_{12} \quad (23)$$

we have

$$\langle \tilde{E}(1) \tilde{E}(2) \rangle = G(1, 1'; 2, 2') : \langle \tilde{J}_t(1') \tilde{J}_t(2') \rangle \quad (24)$$

When the effect of  $I_{12}$  is small it is useful to define renormalized, collective modes or wave types by

$$(Y_M + P) \cdot \tilde{E}_\alpha = 0, \quad \alpha = 1, 2, \dots \quad (25)$$

The derivation of kinetic and Klimontovich equations for interacting quasi-particle now follows in a straightforward way from previous results.<sup>3</sup>

Office of Naval Research  
N00014-76-C-0176

S. Barone and N. Marcuvitz

National Science Foundation  
ENG-76-21829

#### REFERENCES

1. J. Weinstock, "Formulation of a Statistical Theory of Strong Plasma Turbulence," *Phys. Fluids* **12**, 1045 (1969); "Turbulent Plasmas in a Magnetic Field - A Statistical Theory," *Phys. Fluids* **13**, 2308 (1970).
2. J. Schwinger, "On the Green's Functions of Quantized Fields I, II," *Proc. Nat. Acad. Sci.* **37**, 452 and 455 (1951).
3. N. Marcuvitz, "On the Theory of Plasma Turbulence," *J. Math. Phys.* **15**, 870 (1974); S. Barone and N. Marcuvitz, "Turbulence in Plasma-Like Systems," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77 (1977); S. Barone, "On the Theory of Plasma Turbulence," preprint.

## A NEW APPROACH TO SOME NONLINEAR IONOSPHERIC PLASMA TURBULENCE PROBLEMS

S. R. Barone

Plasma phenomena in which electron and ion inertia play a minor role are often modeled by special cases of the two fluid equations<sup>1-3</sup>

$$\frac{\partial n}{\partial t} + \nabla \cdot (n \underline{v}_e) = 0, \quad (1)$$

$$\frac{\partial n}{\partial t} + \nabla \cdot (n \underline{v}_i) = 0, \quad (2)$$

$$n e (\underline{E} + \underline{v}_e \times \underline{B}) + \nabla P_e + n m_e \nu_{en} \underline{v}_e + n m_e \nu_{ei} (\underline{v}_e - \underline{v}_i) - n m_e \underline{g} = 0, \quad (3)$$

$$-n e (\underline{E} + \underline{v}_i \times \underline{B}) + \nabla P_i + n m_i \nu_{in} \underline{v}_i + n m_i \nu_{ie} (\underline{v}_i - \underline{v}_e) - n m_i \underline{g} = 0, \quad (4)$$

$$\nabla \times \underline{E} = 0, \quad (5)$$

in which both electron and ion inertia have been neglected; the electron and ion densities have been set equal; the transverse part of the electromagnetic field is omitted;  $\underline{B}$  is a uniform, static magnetic field;  $\nu_{en}$ ,  $\nu_{in}$  are the electron and ion collision frequencies with neutrals,  $\nu_{ei}$  and  $\nu_{ie}$  are the electron and ion collision frequencies with each other (all assumed to be independent of position) and  $\underline{g}$  is the acceleration due to gravity. These equations are written in the frame of the neutrals and are ordinarily closed by assuming either adiabatic or isothermal behavior for the electrons and ions.

The purpose of this report is to point out that for two-dimensional flows and/or turbulence perpendicular to  $\underline{B}$  there is a useful way to introduce scalar and vector potentials. For simplicity of presentation we suppose that  $P_e = n k_B T_e$ ,  $P_i = n k_B T_i$ , where  $k_B$  is Boltzmann's constant and the electron and ion temperatures,  $T_e$ ,  $T_i$  are constant. Set

$$n \underline{v}_e = \nabla \chi + \frac{m_i (\nu_{in} + \nu_{ie}) - m_e \nu_{ei}}{m_e \nu_{en} + m_i \nu_{in}} \nabla \chi \underline{a} + \frac{m_e + m_i}{m_e \nu_{en} + m_i \nu_{in}} n \underline{g}, \quad (6)$$

$$n \underline{v}_i = \nabla \chi - \frac{m_e (\nu_{en} + \nu_{ei}) - m_i \nu_{ie}}{m_e \nu_{en} + m_i \nu_{in}} \nabla \chi \underline{a} + \frac{m_e + m_i}{m_e \nu_{en} + m_i \nu_{in}} n \underline{g}, \quad (7)$$

where  $\chi$  is a scalar potential and the vector potential  $\underline{a}$  has only a single component parallel to  $\underline{B}$ . Substituting into Eqs. (1) to (4) yields:

1) The component of  $\underline{a}$  in terms of  $n$  and  $\chi$ ,

$$e \underline{B} \cdot \underline{a} = k_B (T_e + T_i) n + (m_e \nu_{en} + m_i \nu_{in}) \chi, \quad (8)$$

2) The electric field in terms of  $n$  and  $\chi$ ,

$$\begin{aligned}
 -en \underline{E} = & \left[ \frac{m_e(\nu_{en} + \nu_{ei}) - m_i \nu_{ie}}{m_e \nu_{en} + m_i \nu_{in}} k_B T_e - \frac{m_i(\nu_{in} + \nu_{ie}) - m_e \nu_{ei}}{m_e \nu_{en} + m_i \nu_{in}} k_B T_i \right] \nabla n \\
 & + [m_e(\nu_{en} + \nu_{ei}) - m_i(\nu_{in} + \nu_{ie})] \nabla \chi \\
 & + \frac{(\nu_{en} \nu_{in} + \nu_{en} \nu_{ie} + \nu_{in} \nu_{ei}) / \Omega_e \Omega_i}{m_e \nu_{en} + m_i \nu_{in}} k_B (T_e + T_i) \nabla n \times e \underline{B} \\
 & + \left( 1 + \frac{\nu_{en} \nu_{in} + \nu_{en} \nu_{ie} + \nu_{in} \nu_{ei}}{\Omega_e \Omega_i} \right) \nabla \chi \times e \underline{B} + \frac{m_e m_i (\nu_{en} - \nu_{in})}{m_e \nu_{en} + m_i \nu_{in}} n \underline{g} \\
 & + \frac{m_e + m_i}{m_e \nu_{en} + m_i \nu_{in}} n \underline{g} \times e \underline{B}, \quad (9)
 \end{aligned}$$

where  $\Omega_{e,i} \equiv eB/m_{e,i}$  are the electron and ion cyclotron frequencies, and

3) A connection between  $n$  and  $\chi$  viz.,

$$\frac{\partial n}{\partial t} + \underline{v}_g \cdot \nabla n = -\nabla^2 \chi, \quad (10)$$

where

$$\underline{v}_g \equiv \frac{m_e + m_i}{m_e \nu_{en} + m_i \nu_{in}} \underline{g}, \quad (11)$$

and we have assumed that the sources of the gravitational field are outside the region of interest so that  $\nabla \cdot \underline{g} = 0$ .

The final equation (5) can be expressed in a variety of forms. Taking the curl of Eq. (9), we have

$$\begin{aligned}
 & \frac{(\nu_{en} \nu_{in} + \nu_{en} \nu_{ie} + \nu_{in} \nu_{ei}) / \Omega_e \Omega_i}{m_e \nu_{en} + m_i \nu_{in}} k_B (T_e + T_i) \nabla^2 n \\
 & + \left( 1 + \frac{\nu_{en} \nu_{in} + \nu_{en} \nu_{ie} + \nu_{in} \nu_{ei}}{\Omega_e \Omega_i} \right) \nabla^2 \chi \\
 & = \nabla n \cdot \left( \frac{\underline{E} \times \underline{B}}{B^2} - \underline{v}_g + (\nu_{en} - \nu_{in}) \frac{m_e m_i}{m_e + m_i} \frac{\underline{v}_g \times e \underline{B}}{(eB)^2} \right). \quad (12)
 \end{aligned}$$

Using Eq. (9) to eliminate  $\underline{E}$  yields a second relation between  $n$  and  $\chi$ , viz.,

$$\nabla^2 \chi + \kappa \left[ \nabla^2 - \left( \frac{\nabla n}{n} \right)^2 \right] n = \left( \frac{\nabla n}{n} \right) \cdot \nabla \chi, \quad (13)$$

where

$$\kappa \equiv \frac{1}{1 + \frac{\Omega_e \Omega_i}{v_{en} v_{in} + v_{en} v_{ie} + v_{in} v_{ei}}} \frac{k_B (T_e + T_i)}{m_e v_{en} + m_i v_{in}}, \quad (14)$$

$$\bar{\nabla} \equiv \nabla + \lambda \frac{\mathbf{B}}{B} \times \nabla, \quad (15)$$

$$\lambda \equiv \frac{\frac{v_{en} + v_{ei}}{\Omega_e} - \frac{v_{in} + v_{ie}}{\Omega_i}}{1 + \frac{v_{en} v_{in} + v_{en} v_{ie} + v_{in} v_{ei}}{\Omega_e \Omega_i}}, \quad (16)$$

and we have assumed that  $\mathbf{g}$  has no component parallel to  $\mathbf{B}$ . According to the present point of view, one solves the evolution equation (10) simultaneously with the constraint (13) for  $\mathbf{n}$  and  $\chi$  and calculates the velocity vectors and electric field via Equations (6) to (9). Examples of the utility of the technique will be discussed elsewhere.

Office of Naval Research  
N00014-76-C-0176

S.R. Barone

#### REFERENCES

1. R.L. Ferch and R.N. Sudan, "Numerical Simulations of Type II Gradient Drift Irregularities in the Equatorial Electrojet," J. Geophys. Res., 82(16), pp. 2283-2288 (1977).
2. B.E. McDonald, S.L. Ossakow, S.T. Zalesak and N.J. Zabusky, "Determination of Minimum Scale Sizes in Plasma Cloud Striations," presented at the 1978 Symp. on the Effect of Ionosphere on Space and Terrestrial Systems, Washington, D.C. (January 24-26, 1978), preprint.
3. E. Ott, "Theory of Rayleigh-Taylor Bubbles in the Equatorial Ionosphere," J. Geophys. Res., 83(A5), pp. 2066-2070 (1978).

## PRELIMINARY REPORT OF NUMERICAL SIMULATION OF TYPE II IRREGULARITIES IN THE EQUATORIAL ELECTROJET

S. Barone, N. Marcuvitz, R. Pascone and N. Solimene

Type II irregularities in the ionization density in the equatorial E-region are believed to result from the mutual interaction of linearly unstable, long wavelength, horizontally propagating (east-west) waves and shorter wavelength vertically propagating waves which become unstable in the presence of the horizontally propagating waves.<sup>1,2</sup> Numerical studies of the evolution towards the steady-state have been carried out by McDonald<sup>3,4</sup> et al., Sato and Ogawa<sup>5</sup> and Ferch and Sudan.<sup>6</sup> These authors all work with the two fluid equations

$$\frac{\partial n}{\partial t} + \nabla \cdot (n \underline{v}_e) = 0, \quad (1)$$

$$\frac{\partial n}{\partial t} + \nabla \cdot (n \underline{v}_i) = 0, \quad (2)$$

$$+ e(\underline{E} + \underline{v}_e \times \underline{B}_0) + k_B T_e \frac{\nabla n}{n} + m_e \nu_e \underline{v}_e = 0, \quad (3)$$

$$- e(\underline{E} + \underline{v}_i \times \underline{B}_0) + k_B T_i \frac{\nabla n}{n} + m_i \nu_i \underline{v}_i = 0, \quad (4)$$

$$\nabla \times \underline{E} = 0, \quad (5)$$

in which both electron and ion inertia have been neglected, the electron and ion densities have been set equal, the transverse part of the electromagnetic field is omitted,  $\underline{B}_0$  is the magnetic field of the earth and  $\nu_{e,i}$  are the electron and ion collision frequencies with neutrals.

For two-dimensional turbulence perpendicular to  $\underline{B}_0$  this system of equations can, without approximation, be reformulated as the single equation<sup>7</sup>

$$(1 + e^{-\psi} \nabla \psi \cdot \overline{\nabla} g e^{\psi}) \frac{\partial \psi}{\partial t} = \kappa \nabla^2 \psi - e^{-\psi} \underline{v}_0 \cdot \nabla \psi, \quad (6)$$

where  $\psi$  is related to the electron density by

$$n = n_0 e^{\psi}, \quad (7)$$

$n_0$  is constant,  $g$  is the Green's function for the two-dimensional Laplacian, i.e.,

$$-\nabla^2 g(\underline{r}, \underline{r}') = \delta(\underline{r} - \underline{r}'), \quad (8)$$

the differential operator  $\overline{\nabla}$  is a linear combination of the gradient and curl operations

$$\overline{\nabla} \equiv \nabla - \lambda \nabla \times \frac{\underline{B}_0}{B_0}, \quad (9)$$

and the constants  $\kappa$ ,  $\lambda$ ,  $\underline{v}_0$  are determined by the system parameters. For sufficiently weak turbulence one might expect a quadratic approximation to equation (6) to be sufficient. To second order in  $\psi$

$$\left[ \frac{\partial}{\partial t} - \kappa \nabla^2 + (1 - \psi) \underline{v}_0 \cdot \nabla \right] \psi = \nabla \psi \cdot (\underline{D} + \kappa \underline{1}) \cdot \nabla \psi \quad (10)$$

where  $\underline{D} \equiv \nabla \times \underline{v}_0$  and  $\underline{1} \equiv -g \nabla^2$ . The nonlinear term on the left has the effect of changing the parameter  $\underline{v}_0$  somewhat as the level of turbulence increases. The relative importance of the nonlinear  $\kappa$  and  $\underline{D}$  terms on the right depends on the numerical values of the parameters and the wavelength under consideration.

For the present situation suppose that the background electron density increases exponentially with altitude (z-coordinate) i. e.,

$$\psi = \frac{z}{L} + \psi' \quad (11)$$

where  $L$  is the vertical scale, and that the electron drift velocity is predominantly towards the west (y-coordinate). Then without approximation

$$\begin{aligned} & \left[ \frac{\partial}{\partial t} - \kappa \nabla^2 + (1 - \psi) \underline{v}_0 \cdot \nabla - \frac{1}{L} \underline{z}^0 \cdot (\underline{D} + \kappa \underline{1}) \cdot \nabla \right] \psi' \\ & = \nabla \psi' \cdot (\underline{D} + \kappa \underline{1}) \cdot \nabla \psi' \quad (12) \end{aligned}$$

where  $\underline{z}^0$  is a unit vector in the (vertical) z-direction. The linear dispersion relation for waves of the form

$$\psi' \propto e^{i \underline{k} \cdot \underline{r} - i \omega t} \quad (13)$$

can be read directly off Eq. (12) viz.

$$\omega = \left( \underline{k} \cdot \underline{v}_0 - \frac{\kappa}{L} k_z \right) - i \left( \kappa k^2 + \frac{1}{L} \frac{\underline{k} \cdot \underline{v}_0 \cdot \underline{k}}{k^2} \right) \quad (14)$$

where  $\underline{\bar{k}} = \underline{k} + \lambda \underline{k} \times \underline{z}^0$  is a unit vector in the direction of the earth's magnetic field. In the E-region electron-ion collisions are negligible and

$$\kappa \equiv \frac{\frac{\nu_e \nu_i}{\Omega_e \Omega_i}}{1 + \frac{\nu_e \nu_i}{\Omega_e \Omega_i}} \frac{k_B (T_e + T_i)}{m_e \nu_e + m_i \nu_i} \quad (15)$$

$$\lambda \equiv \frac{\frac{v_i}{\Omega_i} - \frac{v_e}{\Omega_e}}{1 + \frac{v_e v_i}{\Omega_e \Omega_i}} \quad (16)$$

Numerical values appropriate to the daytime equatorial electrojet are:

$$v_e = 4 \times 10^4 / \text{sec.},$$

$$v_i = 2.5 \times 10^3 / \text{sec.},$$

$$\Omega_e = 5 \times 10^6 / \text{sec.},$$

$$\Omega_i = 90 / \text{sec.},$$

$$C_s^2 \equiv k_B (T_e + T_i / m_i) = 10^5 \text{ m}^2 / \text{sec.}^2,$$

$$L = 6000 \text{ m},$$

$$V_d = 100 \text{ m/sec.},$$

where  $V_d$  is the electron drift velocity. Thus,

$$\kappa = 7.3 \text{ m}^2 / \text{sec.},$$

$$\lambda = 23,$$

and the electron drift velocity  $V_d$  and the velocity parameter  $v_o$ , are approximately related by

$$v_o \approx \frac{V_d}{1 + \frac{v_e v_i}{\Omega_e \Omega_i}} = 82 \text{ m/sec.} \quad (17)$$

Under these circumstances the  $1/L$  contribution to the real part of  $\omega$  is negligible,  $\bar{k} \approx \lambda k \times \hat{x}^0$  and Eq. (14) becomes

$$\omega = k \cdot v_o - i \left( \kappa k^2 - \frac{\lambda}{L} \frac{k \cdot v_o}{k^2} k_y \right). \quad (18)$$

Waves traveling towards the west are unstable when

$$k^4 < k_c^2 k_y^2, \quad (19)$$

where  $k_c \equiv \sqrt{\lambda v_o / L_K} \cong 2\pi / 30$  m. Alternatively, linear instability occurs when  $k_y < k_c$  for waves which travel at an angle of elevation ( $\theta$ ) small enough that

$$\cos \theta > \sqrt{k_y / k_c} \quad (20)$$

Run 4, (Figs. 4-1(a)(b), 6(a)(b), 11(a)(b)), is for an initial distribution

$$\psi(t=0) = .50 \cos \frac{2\pi y}{100} \cos \frac{2\pi z}{100}$$

where  $y, z$  are measured in meters. According to the linearized theory this wave will grow exponentially ( $\exp \gamma t$ ) with  $\gamma \cong 0.10/\text{sec}$ . Actually this initial distribution grows at a somewhat smaller rate for about 0.3 sec. at which time significant distortion is apparent and the maximum value of the distribution begins to decrease. The maximum value of the distribution increases again at 0.5 sec. and decreases monotonically thereafter, i.e., until  $t=1.2$  sec. Throughout this time interval the wavenumber spectrum spreads considerably in  $z$  corresponding to the development of sharp variations in the  $z$ -direction. There is no evidence of any turbulent structure developing.

Run 9, (Figs. 9-1(a)(b), 4(a)(b), 11(a)(b), 18(a)(b), 24(a)(b)) is for an initial distribution

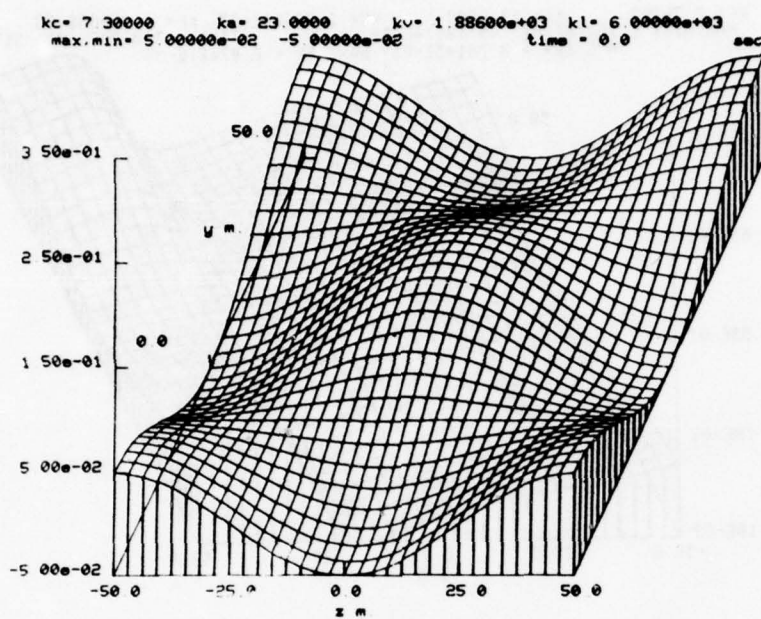
$$\psi(t=0) = -.03 \sin \frac{2\pi y}{72} - 0.1 \sin \frac{2\pi z}{72}$$

where again  $y, z$  are measured in meters. The linear growth rate for the wave traveling in the  $y$ -direction is  $\gamma \cong .26/\text{sec}$ . while according to the linear theory the standing wave in the  $z$ -direction decays exponentially with  $\gamma \cong .056/\text{sec}$ . ( $\cong 18$  sec. decay time). For the time duration of this run (2.3 sec.) the relatively large wave traveling in the  $y$ -direction simply grows exponentially with the linear growth rate as if the relatively small standing wave in the  $z$ -direction was not present. The relatively small wave however does not evolve linearly. In the presence of the  $y$  directed wave the vertical wave breaks up into three (or more) waves with the same propagation constants in the  $z$ -direction as the original wave but with propagation constants in the  $y$ -direction equal to  $\pm$  the  $y$  propagation constant of the large wave. The amplitudes of these three (or more) components oscillate in time. Again, there is no evidence of turbulent structures developing.

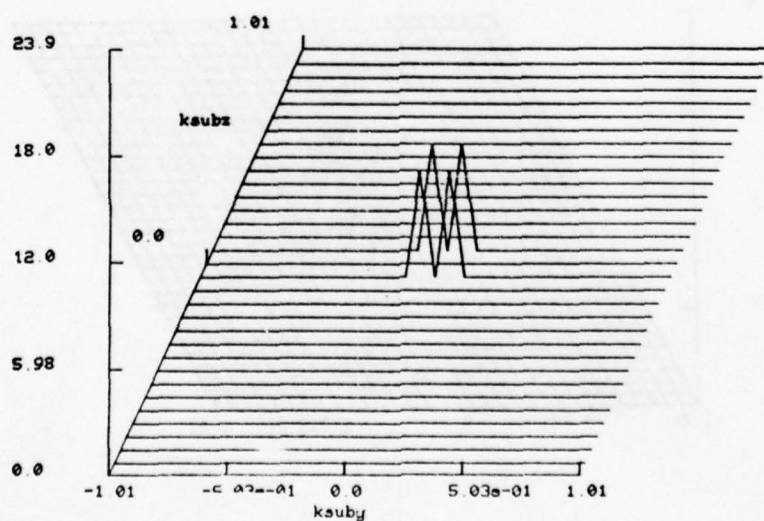
Run 11, (Figs. 11-1 (a)(b), 13 (a)(b), 25 (a)(b), 36 (a)(b)) for an initial distribution

$$\psi(t=0) = .04 \cos \frac{2\pi y}{100} + .03 \cos \frac{2\pi z}{100}$$

This distribution differs from the previous one in that the two waves present are of

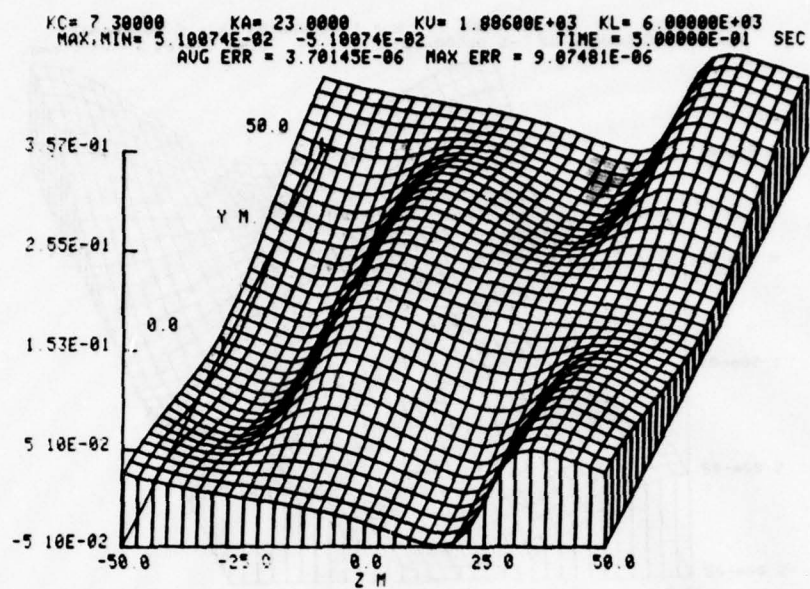


Power Spectrum     $\text{time} = 0.0 \quad \text{sec.}$   
 $k_a = 5.00000e-02$      $k_v = 1.88600e+03$      $k_l = 6.00000e+03$

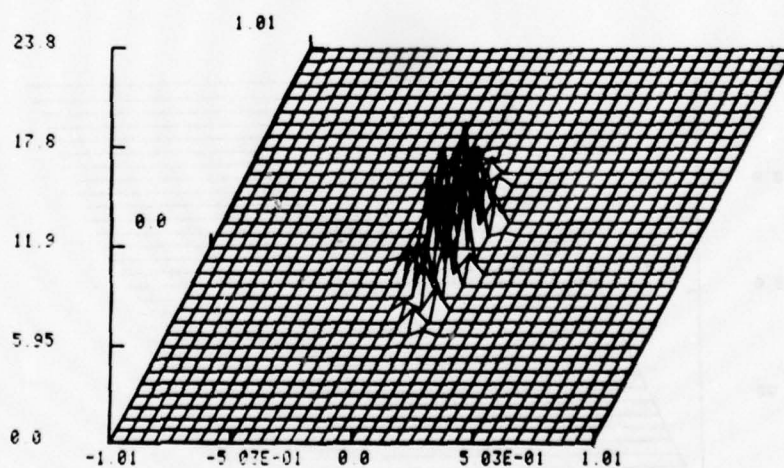


Run 4 (Figs. 4-1(a)(b)).

## WAVE-MATTER INTERACTIONS



$\backslash POWER \backslash SPECTRUM$      $TIME = 5.00000E-01 \text{ SEC}$   
 $KE = 5.00000E-02$      $KU = 1.00600E+03$      $KL = 6.00000E+03$

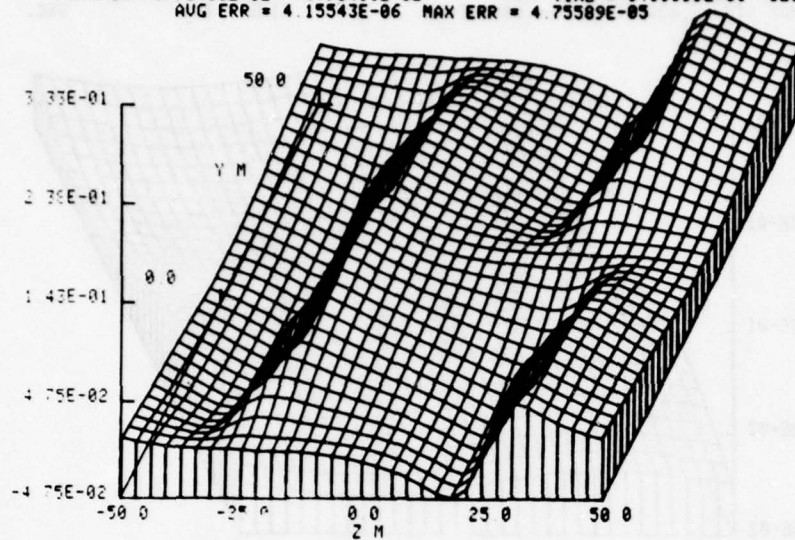


Run 4 (Figs. 4-6(a)(b)).

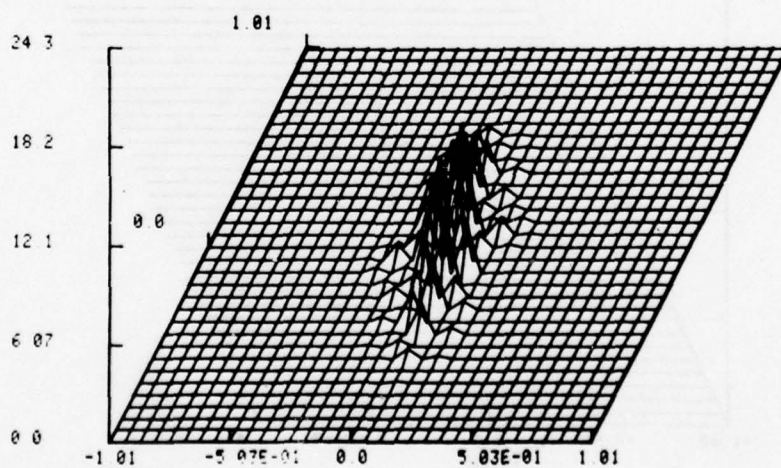
# WAVE-MATTER INTERACTIONS

255

$KC = 7.30000$      $KA = 23.0000$      $KU = 1.88600E+03$      $KL = 6.00000E+03$   
 $MAX MIN = 4.75408E-02$      $-4.75411E-02$      $TIME = 9.99999E-01$  SEC.  
 $AUG ERR = 4.15543E-06$      $MAX ERR = 4.75589E-05$



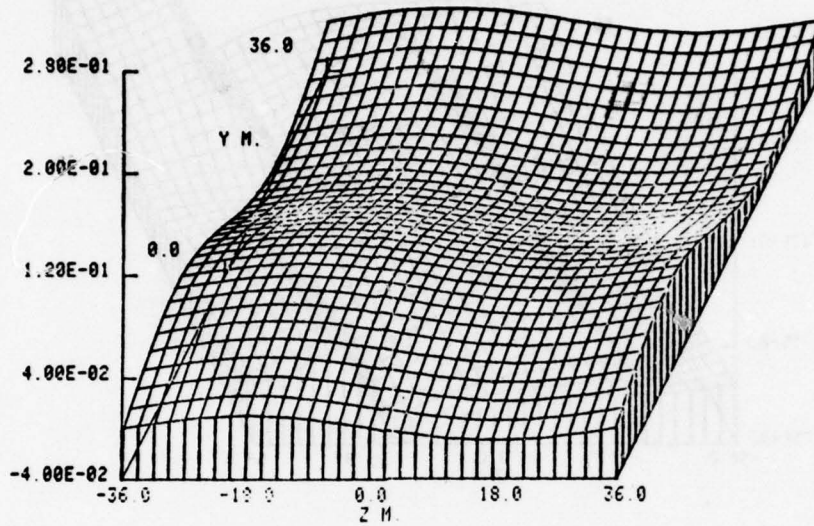
POWER SPECTRUM     $TIME = 9.99999E-01$  SEC.  
 $KE = 5.00000E-02$      $KU = 1.88600E+03$      $KL = 6.00000E+03$



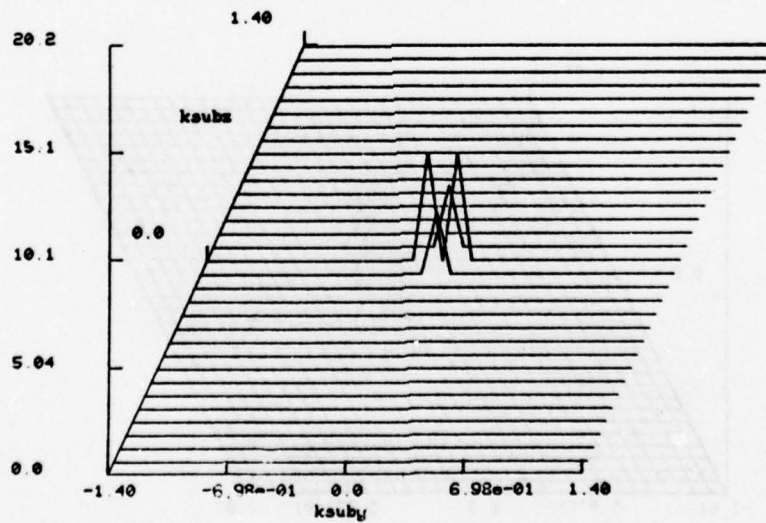
Run 4 (Figs. 4-11(a)(b)).

## WAVE-MATTER INTERACTIONS

KC= 7.30000 KA= 23.0000 KU= 1.88600E+03 KL= 6.00000E+03  
 MAX,MIN= 4.00000E-02 -4.00000E-02 TIME = 0.0 SEC.

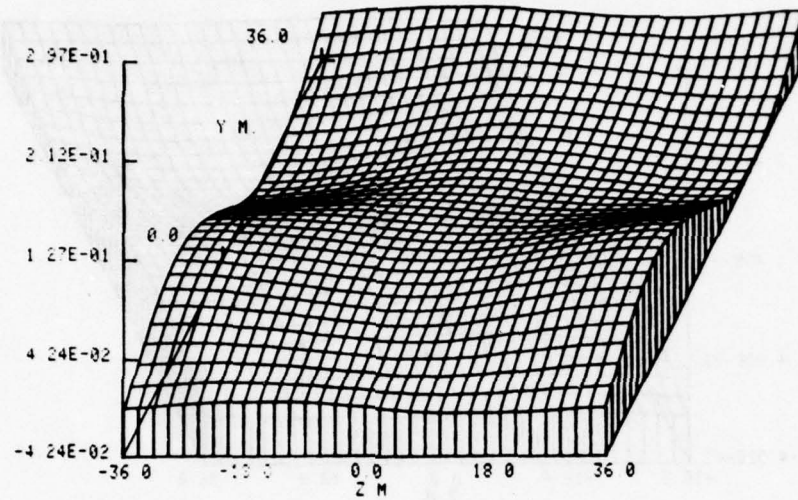


Power Spectrum time = 0.0 sec.  
 $k_x = -3.00000E-02$   $k_y = 1.88600E+03$   $k_z = 6.00000E+03$

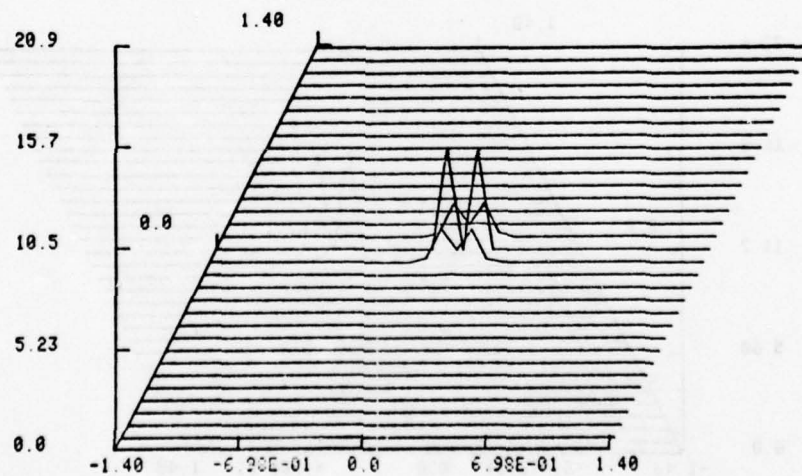


Run 9 (Figs. 9-1(a)(b)).

$\gamma C = 7.30000$      $K_A = 23.0000$      $K_U = 1.88600E+03$      $K_L = 6.00000E+03$   
 $\text{MAX, MIN} = 4.24027E-02 \quad -4.24022E-02$      $\text{TIME} = 3.00000E-01 \text{ SEC}$   
 $\text{AUG ERR} = 5.72648E-06$      $\text{MAX ERR} = 1.65068E-05$



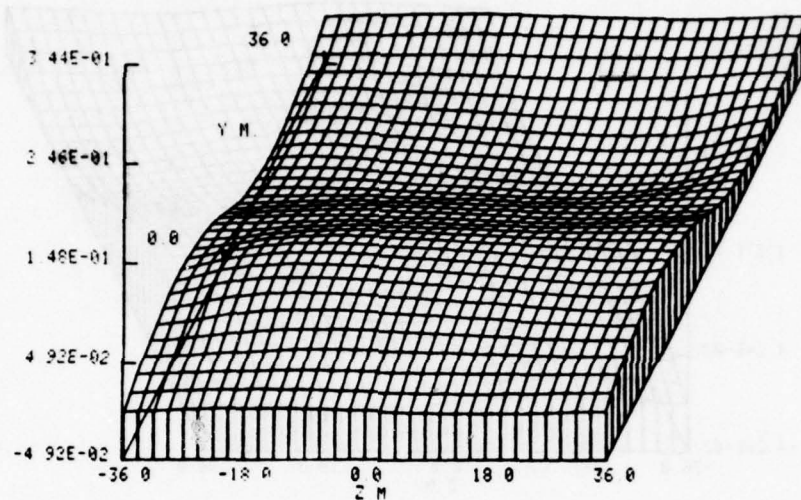
$\backslash \text{POWER} \backslash \text{SPECTRUM}$      $\text{TIME} = 3.00000E-01 \text{ SEC}$   
 $K_E = -3.00000E-02$      $K_U = 1.88600E+03$      $K_L = 6.00000E+03$



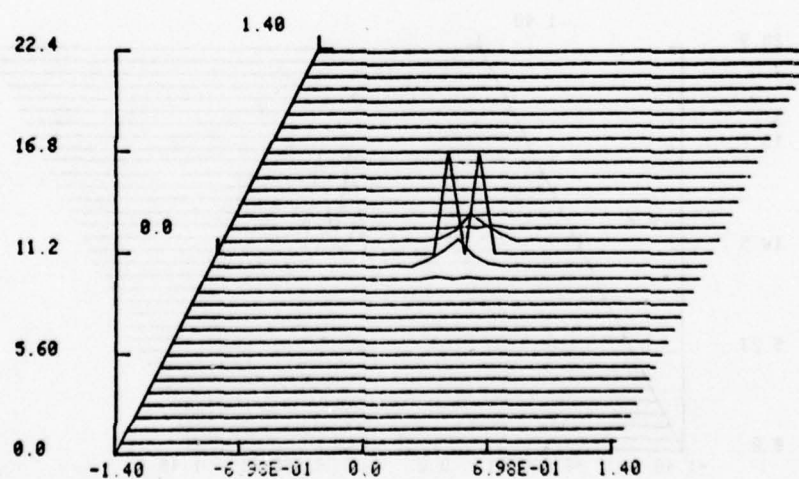
Run 9 (Figs. 9-4(a)(b)).

## WAVE-MATTER INTERACTIONS

$KC = 7.30000$      $KR = 23.0000$      $KU = 1.88600E+03$      $KL = 6.00000E+03$   
 $MAX MIN = 4.92083E-02$      $-4.92105E-02$      $TIME = 9.99999E-01$  SEC  
 $AUG ERR = 9.60706E-06$      $MAX ERR = 6.80983E-05$

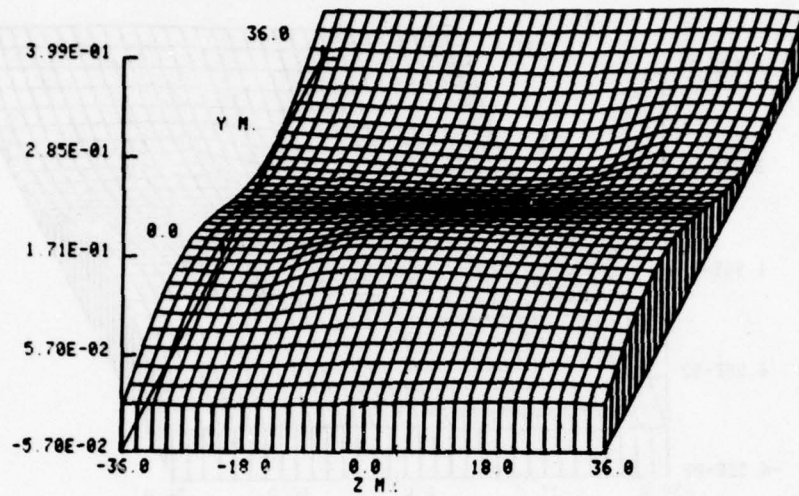


$\backslash POWER \backslash SPECTRUM$      $TIME = 9.99999E-01$  SEC  
 $KE = -3.00000E-02$      $KU = 1.88600E+03$      $KL = 6.00000E+03$

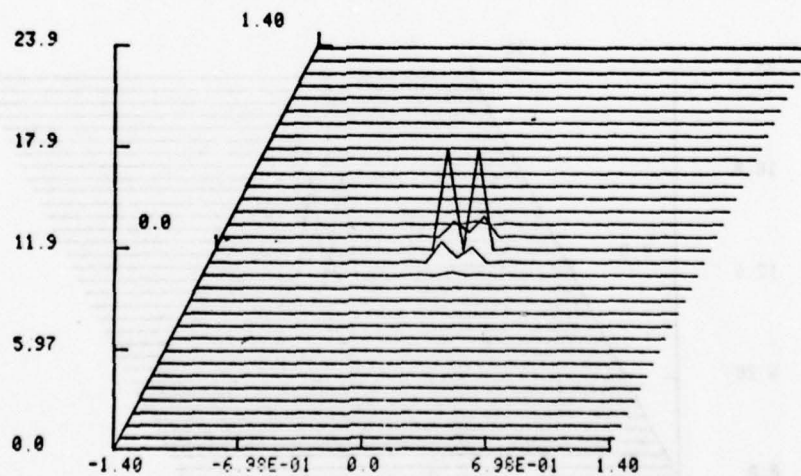


Run 9 (Figs. 9-11(a)(b)).

KC= 7.30000 KA= 23.0000 KU= 1.88600E+03 KL= 6.00000E+03  
 MAX, MIN= 5.69904E-02 -5.69834E-02 TIME = 1.70000 SEC.  
 AVG ERR = 1.36169E-05 MAX ERR = 2.91113E-04



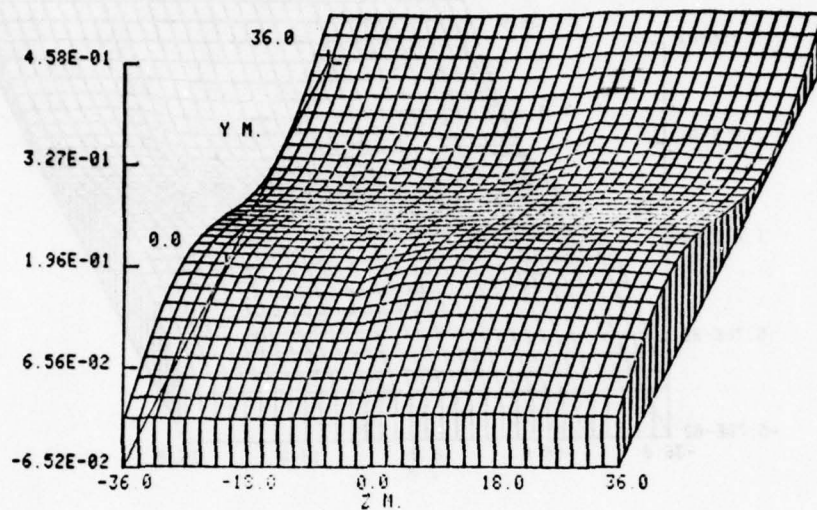
\POWER \SPECTRUM TIME = 1.70000 SEC.  
 KE=-3.00000E-02 KU= 1.88600E+03 KL= 6.00000E+03



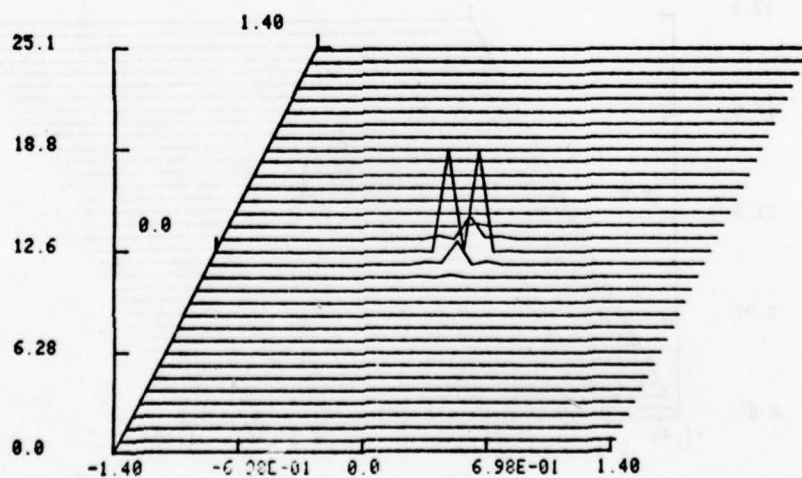
Run 9 (Figs. 9-18(a)(b)).

## WAVE-MATTER INTERACTIONS

$KC = 7.30000$      $KA = 23.0000$      $KU = 1.88600E+03$      $KL = 6.00000E+03$     SEC.  
 $MAX, MIN = 6.56478E-02 \quad -6.51823E-02$      $TIME = 2.30000$   
 $AUG ERR = 9.13036E-05$      $MAX ERR = 1.46366E-03$



$\backslash POWER \backslash SPECTRUM$      $TIME = 2.30000$     SEC.  
 $KE = -3.00000E-02$      $KU = 1.88600E+03$      $KL = 6.00000E+03$

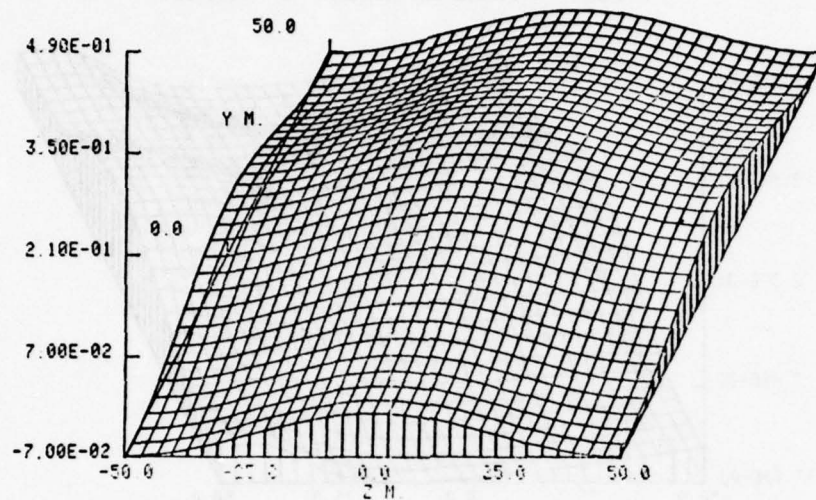


Run 9 (Figs. 9-24(a)(b)).

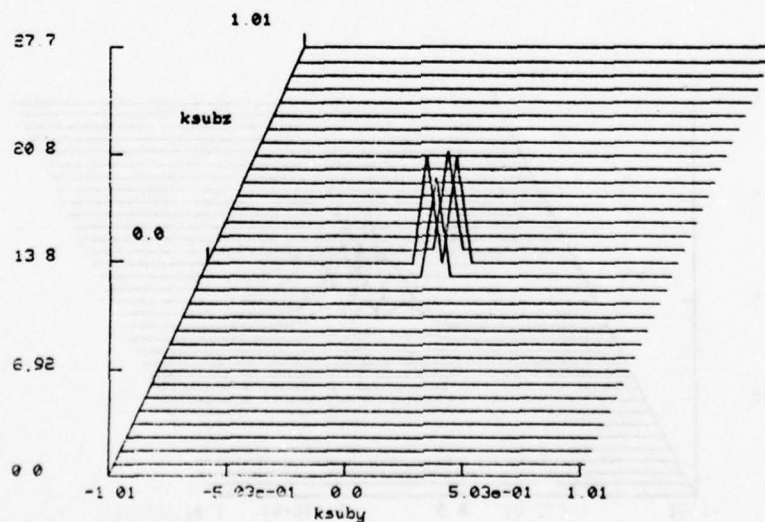
# WAVE-MATTER INTERACTIONS

261

KC= 7.30000 KA= 23.0000 KV= 1.88600E+03 KL= 6.00000E+03  
 MAX,MIN= 7.00000E-02 -7.00000E-02 TIME = 0.0 SEC.



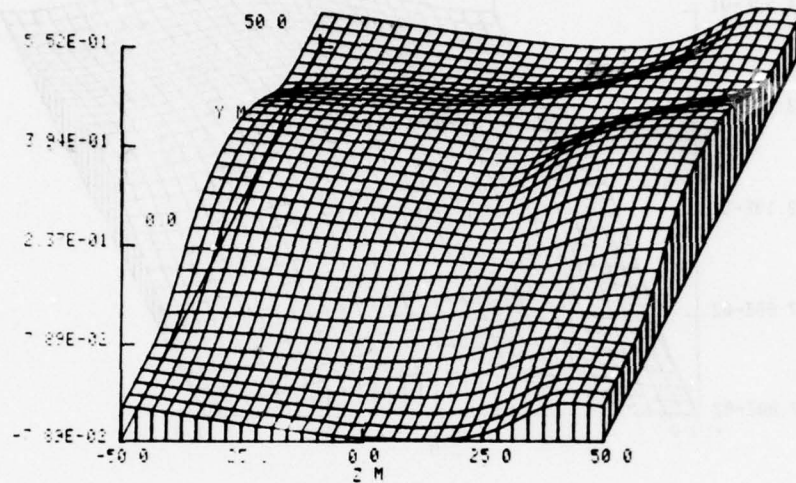
Power Spectrum time = 0.0 sec.  
 $k_x = 4.00000E-02$   $k_y = 1.88600E+03$   $k_z = 6.00000E+03$



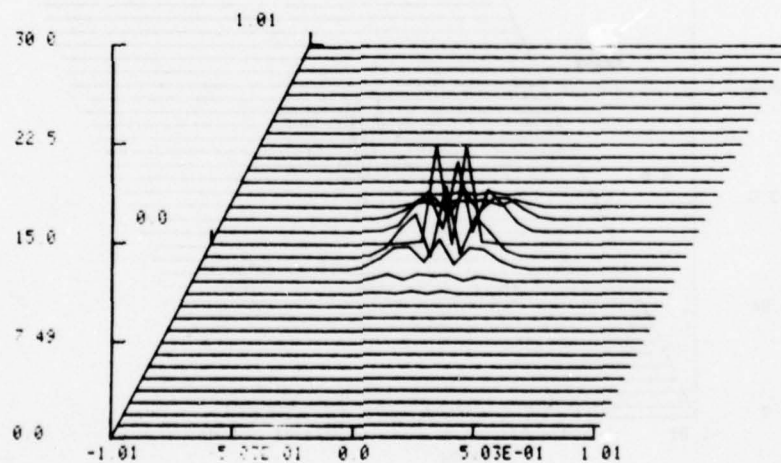
Run 11 (Figs. 11-1(a)(b)).

## WAVE-MATTER INTERACTIONS

$KC = 7.30000$      $KA = 23.0000$      $KU = 1.88600E+03$      $KL = 6.00000E+03$   
 $MAX MIN = 7.89029E-02 \quad -7.88911E-02$      $TIME = 6.00000E-01 \text{ SEC}$   
 $AUG ERR = 8.20630E-06$      $MAX ERR = 8.17515E-05$

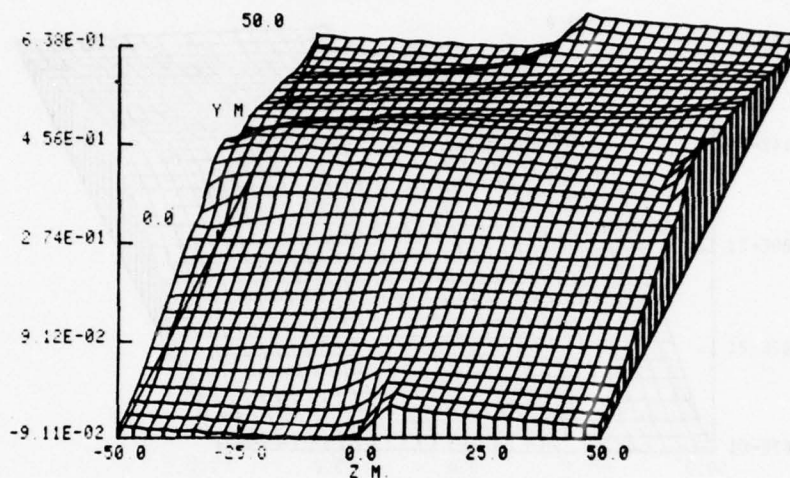


POWER SPECTRUM     $TIME = 6.00000E-01 \text{ SEC}$   
 $KE = 4.00000E-02$      $KU = 1.88600E+03$      $KL = 6.00000E+03$

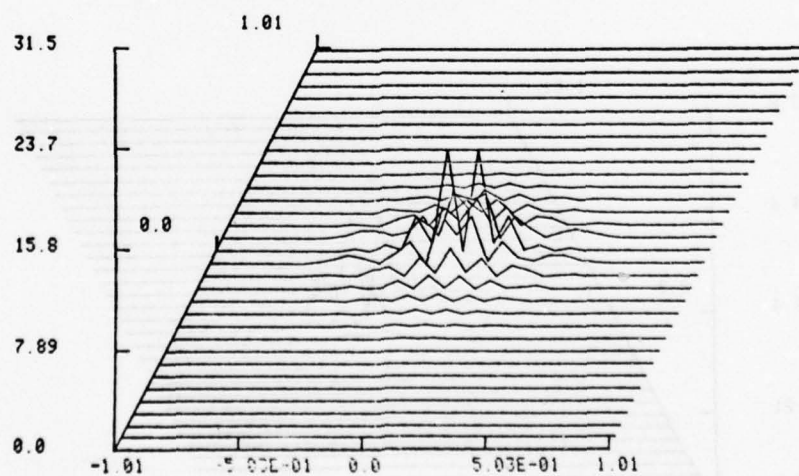


Run 11 (Figs. 11-13(a)(b)).

$KC = 7.30000$      $KA = 23.0000$      $KU = 1.88600E+03$      $KL = 6.00000E+03$   
 $MAX, MIN = 9.11867E-02 \quad -9.11453E-02$      $TIME = 1.20000$     SEC.  
 $AUG ERR = 6.04206E-05$      $MAX ERR = 9.52859E-04$



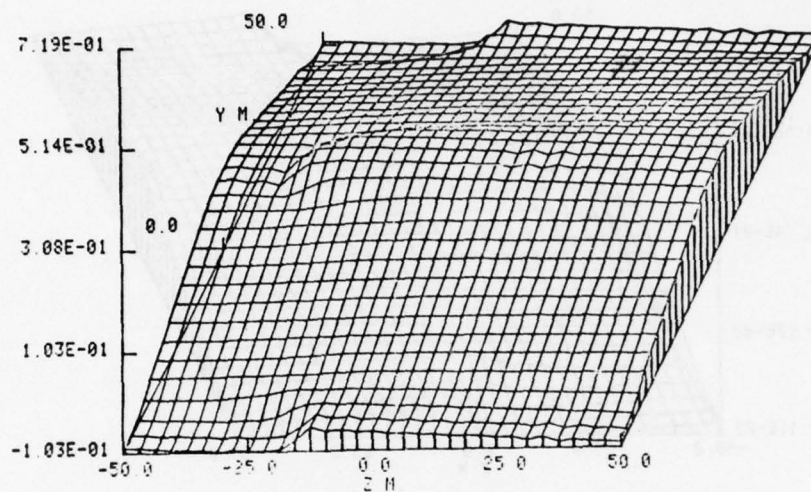
$\backslash POWER \backslash SPECTRUM$      $TIME = 1.20000$     SEC.  
 $KE = 4.00000E-02$      $KU = 1.88600E+03$      $KL = 6.00000E+03$



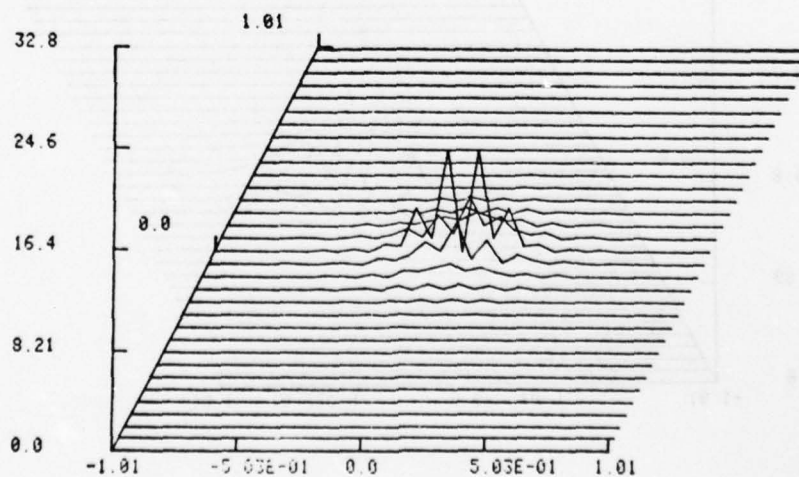
Run 11 (Figs. 11-25(a)(b)).

## WAVE-MATTER INTERACTIONS

$KC = 7.30000$      $KA = 23.0000$      $KV = 1.88600E+03$      $KL = 6.00000E+03$      $TIME = 1.75000$     SEC.  
 $MAX, MIN = 1.02891E-01$      $-1.02590E-01$   
 $AUG ERR = 9.69619E-05$      $MAX ERR = 1.27608E-03$



$\backslash POWER \backslash SPECTRUM$      $TIME = 1.75000$     SEC.  
 $KE = 4.00000E-02$      $KV = 1.88600E+03$      $KL = 6.00000E+03$



Run 11 (Figs. 11-36(a)(b)).

more nearly equal amplitude. At first the y-directed wave grows at a rate approximately double the linear growth rate, i.e., the presence of the standing wave in the z-direction enhances the growth rate for the y-directed wave. Clearly, the reason for this is that the vertical electron density gradient associated with the z wave exceeds that associated with the vertical scale,  $L=6000$  m. After about .5 sec. the amplitude of the y-directed wave dominates all others and this amplitude grows exponentially with the linear growth rate for the remainder of this run (1.75 sec.). In contrast to the previous run the original standing wave in the z-direction now leads to a spread in the wavenumber spectrum in both the y and z directions. The amplitude spectrum eventually becomes rather complicated. However, because of definite phase relations between the various Fourier components, the electron density in configuration space remains fairly smooth and non-turbulent.

Office of Naval Research  
N00014-76-C-0176

S. Barone

#### REFERENCES

1. Farley, D. T. and B. B. Balsley, "Instabilities in the Equatorial Electrojet," J. Geophys. Res., 78 (1), 227-239 (1973).
2. Sudan, R. N., J. Akinrimisi and D. T. Farley, "Generation of Small Scale Irregularities in the Equatorial Electrojet," J. Geophys. Res., 78 (1), 240-248 (1973).
3. McDonald, B. E., T. P. Coffey, S. Ossakow and R. N. Sudan, "Preliminary Report of Numerical Simulation of Type 2 Irregularities in the Equatorial Electrojet," J. Geophys. Res., 79 (16), 2551-2554 (1974).
4. McDonald, B. E., T. P. Coffey, S. Ossakow and R. N. Sudan, "Numerical Studies of Type 2 Equatorial Electrojet Irregularity Development," Radio Sc., 10 (3), 247-254 (1975).
5. Sato, T., and T. Ogawa, "Self Consistent Studies of Two Dimensional Large Scale ( $\sim 100$  m) Electrojet Irregularities," J. Geophys. Res., 81 (19), 3248-3256 (1976).
6. Ferch, R. L. and R. N. Sudan, "Numerical Simulations of Type II Gradient Drift Irregularities in the Equatorial Electrojet," J. Geophys. Res., 82 2283-2288 (1977).
7. See the preceding article; Barone, S., "A new approach to some nonlinear plasma turbulence problems".

## NONLINEAR THEORY OF TYPE II IRREGULARITIES IN THE EQUATORIAL ELECTROJET

S. Barone

Farley and Balsley<sup>1</sup> and Sudan et. al.<sup>2</sup> have proposed that type II irregularities in the ionization density in the equatorial E-region are to be pictured as a quasi-steady state of two-dimensional turbulence perpendicular to the earth's magnetic field. In the presence of the electrojet current a non-turbulent, smooth increase of ionization density with altitude is linearly unstable for long wavelength horizontally propagating waves. This, gradient drift, instability gives rise to horizontal density gradients and vertical electron velocities, a configuration which is unstable for shorter wavelength, vertically propagating waves. The mutual interaction of horizontally and vertically propagating waves leads to a turbulent steady state in which energy delivered to long wavelengths cascades nonlinearly to shorter wavelengths and eventually returns to heat. The steady state has been studied analytically by Rognlien, and Weinstock<sup>3</sup> and Sudan and Keskinen<sup>4</sup> and numerically by Ferch and Sudan,<sup>5</sup> McDonald et. al.<sup>6,7</sup> and Sato and Ogawa.<sup>8</sup>

In this work the two fluid equations of motion describing type II irregularities in the equatorial electrojet are reformulated in terms of a scalar and vector potential for the fluid momentum densities. Elimination of the vector potential yields a single equation, first order in time, for the time evolution of the ionization density,<sup>9</sup> viz.

$$\left(1 + e^{-\psi} \nabla \psi \cdot \bar{\nabla} g e^{\psi}\right) \frac{\partial \psi}{\partial t} = \kappa \nabla^2 \psi - e^{-\psi} \underline{y}_0 \cdot \nabla \psi \quad (1)$$

Where  $g$  is the Green's function for Poisson's equation and

$$\bar{\nabla} \equiv \nabla + a \frac{\underline{B}}{B} \times \nabla \quad (2)$$

The numerical values of  $\kappa$ ,  $\lambda$  and the constant vector  $\underline{y}_0$  are discussed in the preceding article.

Sudan and Keskinen<sup>4</sup> have given a direct interaction<sup>10-13</sup> analysis of an equation formally the same as the first iteration approximation to Equation (1). Clearly the iterative approximation to Eq. (1) fails for  $\lambda \psi \gtrsim 1$  or, for ionospheric parameters,  $\psi \gtrsim 4\%$ . Since interesting levels of turbulence may exceed 4% we have carried out a direct interaction analysis of the full Equation (1). Numerical results are given for the power law wavenumber spectrum which results from theoretical considerations,<sup>14</sup> computer simulations<sup>5-7</sup> and observations<sup>15</sup> of type II irregularities.

Anticipating that the saturated level of  $\psi$  will be small compared to unity we neglect the three exponential factors in Equation (1). The Fourier - Laplace transform satisfies

$$(\omega - \omega_{\underline{k}}) \psi'(\underline{k}, \omega) = \int \frac{(d\underline{k}')}{(2\pi)^2} \frac{d\omega'}{2\pi} \psi'(\underline{k} - \underline{k}', \omega - \omega') V_{\underline{k}\omega, \underline{k}'\omega'} \psi'(\underline{k}', \omega') , \quad (3)$$

where

$$V_{\underline{k}\omega, \underline{k}'\omega'} \equiv \frac{(\underline{k} - \underline{k}') \cdot \bar{\underline{k}}' \omega' g(\underline{k}')}{1 + i \frac{1}{L} \underline{z}^0 \cdot \bar{\underline{k}} g(\underline{k})} . \quad (4)$$

and  $\omega_{\underline{k}}$  are the small amplitude mode frequencies. The direct interaction approximation as developed by Kraichnan<sup>10-13</sup> leads to a pair of coupled equations for the nonlinear frequency shifts,  $\Gamma_{\underline{k}\omega}$  and two point correlation function,  $I_{\underline{k}\omega}$  viz.,

$$-\Gamma_{\underline{k}\omega} = \int (d\underline{k}') d\omega' \frac{w_{\underline{k}\omega, \underline{k}-\underline{k}'\omega' + \omega'} w_{\underline{k}-\underline{k}'\omega' - \omega', \underline{k}\omega}}{\omega - \omega' - \omega_{\underline{k}-\underline{k}'} + \Gamma_{\underline{k}-\underline{k}'\omega' - \omega'}} I_{\underline{k}'\omega'} , \quad (5)$$

and

$$\left| \omega - \omega_{\underline{k}} + \Gamma_{\underline{k}\omega} \right|^2 I_{\underline{k}\omega} = \frac{1}{2} \int (d\underline{k}') d\omega' \left| w_{\underline{k}\omega, \underline{k}'\omega'} \right|^2 I_{\underline{k}'\omega'} I_{\underline{k}-\underline{k}', \omega - \omega'} , \quad (6)$$

where

$$w_{\underline{k}\omega, \underline{k}'\omega'} \equiv V_{\underline{k}\omega, \underline{k}'\omega'} + V_{\underline{k}\omega, \underline{k}-\underline{k}'\omega' - \omega'} . \quad (7)$$

The angular average of  $\Gamma_{\underline{k}\omega}$  on resonance ( $\omega = \omega_{\underline{k}}$ ) can be shown to satisfy

$$\begin{aligned} -\Gamma_{\underline{k}} &= \int_{\alpha k}^{\infty} k' dk' I_{\underline{k}'} \frac{1}{\Gamma_{\underline{k}'}} \bar{B}_{\underline{k}\underline{k}'} \frac{i}{\sqrt{2}} Z \\ &+ \int_{\alpha k}^{\infty} k' dk' I_{\underline{k}'} \Gamma_{\underline{k}'} \bar{A}_{\underline{k}\underline{k}'} \frac{Z'}{2} \end{aligned} \quad (8)$$

where  $Z$  and  $Z'$  are constants and  $\bar{A}_{\underline{k}\underline{k}'}$ ,  $\bar{B}_{\underline{k}\underline{k}'}$  are known kernels. The solution is

$$\Gamma_k^2 = .66 \left( \frac{3\pi}{8} \right) (a v_o k^2)^2 e^{-2\pi a^2 Z' \int_k^\infty k dk I_k} + 2\pi a^2 Z' \int_k^\infty k dk I_k \int_k^\infty \frac{dk}{k} I_k e^{\dots} \quad (9)$$

When the exponents in the above equation are sufficiently small that they can be neglected i. e., for sufficiently low levels of turbulence or for  $k \rightarrow \infty$

$$|\Gamma_k| \rightarrow \frac{1.2}{\sqrt{n}} a v_o k^2 I_k^{1/2} \quad (10)$$

where  $n$  is the spectral index. Except for the numerical coefficient this is the result of Reference (4). (The numerical value of the plasma dispersion function used in Ref. (4) is apparently incorrect.) The above result is valid for sufficiently low levels of turbulence or wavenumbers large enough that  $2\pi a^2 |Z'| k^2 I_k / (n-2) < 1$  or equivalently  $|\Gamma_k| < k v_o$ . For decreasing wavenumbers or increasing levels of turbulence the exponentials in Eq. (10) are important and the nonlinear modal widths,  $|\Gamma_k|$  increase to the point where they are comparable to the resonant frequencies,  $\omega_k \approx k v_o$ . In particular, for sufficiently high levels of turbulence or for  $k \rightarrow 0$

$$|\Gamma_k| \rightarrow \frac{k v_o}{2 \sqrt{Z'}} \quad , \quad Z' > 0 \quad (11)$$

It is convenient to define the intensity dependent wavenumber,  $k_o$ , by

$$2\pi a^2 |Z'| k^2 I_k = (n-2) \left( \frac{k_o}{k} \right)^{n-2} \quad (12)$$

For any level of turbulence, modes with wavenumbers much larger than  $k_o$  behave according to Eq. (10). Modes with wavenumbers much less than  $k_o$  behave according to Eq. (11). These results are displayed in Figure (1). Notice that  $|\Gamma_k|/k v_o$  is insensitive to both the value of  $n$  ( $3 < n < 4$ ) and  $k$  over the indicated range ( $0 < k \lesssim 1.5 k_o$ ). For  $k$  larger than  $k_o$  the asymptotic result of Eq. (10) (Sudan and Keskinen)<sup>4</sup> is accurate. For  $k \ll k_o$  the present calculation indicates that  $|\Gamma_k|/k v_o$  saturates rather than continuing to increase as  $k$  decreases (presumably  $Z' > 0$ ).

The relationship between the wavenumber  $k_o$  that divides the strong turbulence regime ( $k < k_o$ ) from the relatively weak turbulence regime ( $k > k_o$ ) and the over-all level of turbulence is given by

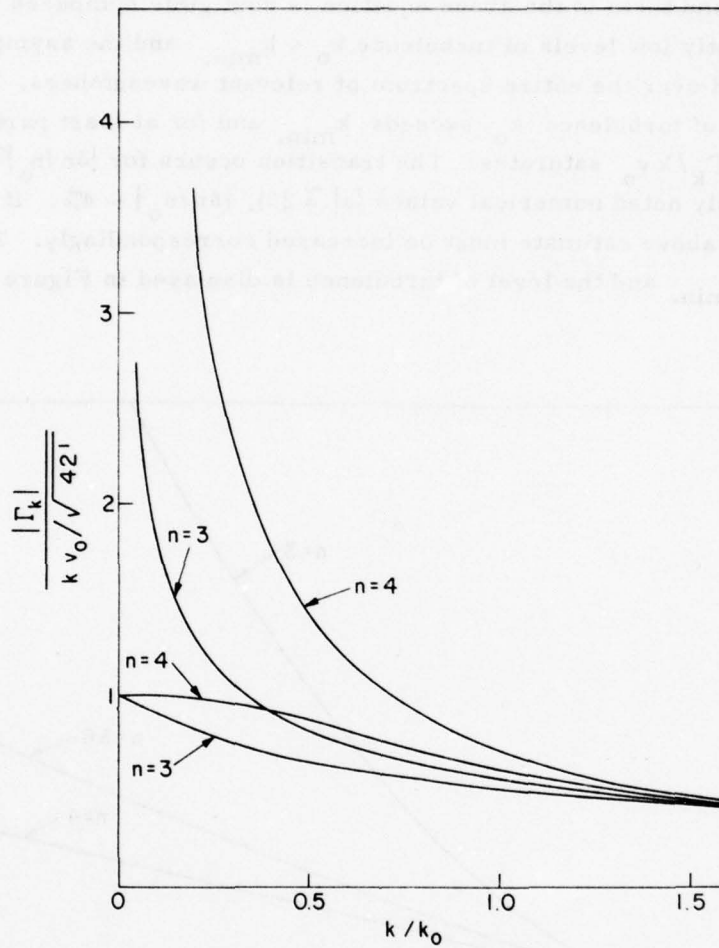


Fig. 1. Nonlinear modal widths,  $\Gamma_k$  as a function of  $k/k_0$ .

$$\left\langle \left| \frac{\delta n}{n_0} \right|^2 \right\rangle = 2\pi \int_{k_{\min.}}^{k_{\max.}} k dk I_k, \quad (13)$$

where  $k_{\min.}$ ,  $k_{\max.}$  are the minimum and maximum wavenumbers present in the turbulent spectrum. Using Eq. (12) to eliminate  $I_k$  in favor of  $k_0$  we have ( $n > 2$ )

$$\left\langle \left| \frac{\delta n}{n_0} \right|^2 \right\rangle = \frac{1}{|Z'| a^2} \left[ \left( \frac{k_0}{k_{\min.}} \right)^{n-2} - \left( \frac{k_0}{k_{\max.}} \right)^{n-2} \right]. \quad (14)$$

Ordinarily the second term in the above equation is negligible compared to the first. Thus, for sufficiently low levels of turbulence  $k_o < k_{min}$ , and the asymptotic result of Eq. (60) is valid over the entire spectrum of relevant wavenumbers. For sufficiently high levels of turbulence  $k_o$  exceeds  $k_{min}$ , and for at least part of the wave-number spectrum  $\Gamma_k/kv_o$  saturates. The transition occurs for  $|\delta n/n_o| \approx 1/\sqrt{|Z'|}a$ , or for the previously noted numerical values  $|a| \approx 23$ ,  $|\delta n/n_o| \sim 4\%$ . If  $|Z'|$  is much less than unity the above estimate must be increased correspondingly. The relationship between  $k_o/k_{min}$  and the level of turbulence is displayed in Figure (2.)

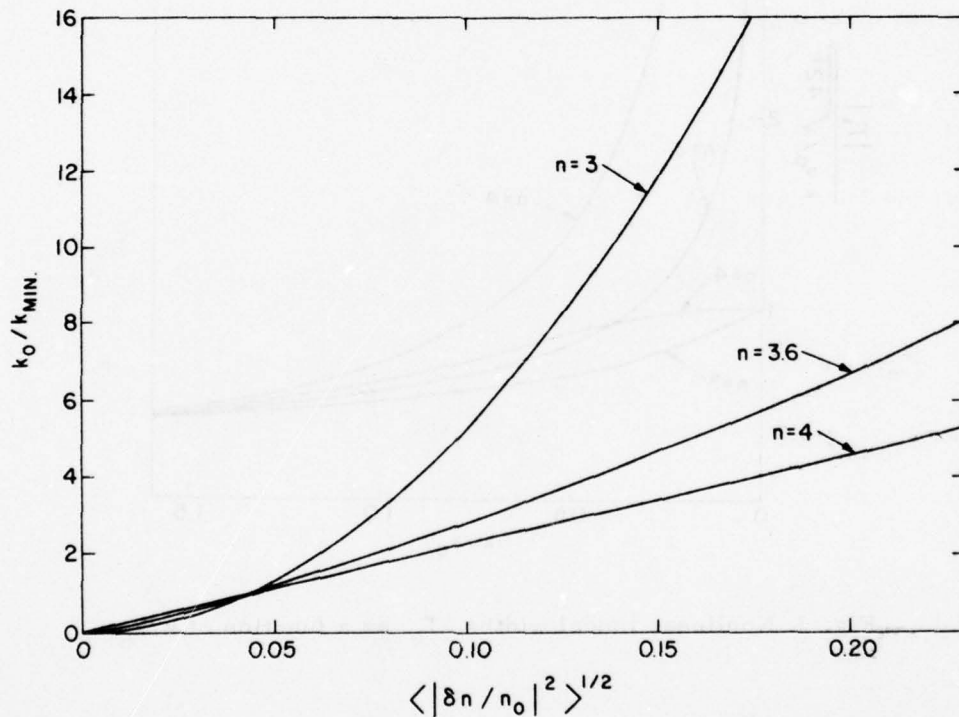


Fig. 2.  $k_o/k_{min}$  as a function of the root-mean-square level of density fluctuations for  $Z'=1$ ,  $|a| = 23$  and various values of the spectral index  $n$ .

## REFERENCES

1. Farley, D.T. and B.B. Balsley, "Instabilities in the Equatorial Electrojet," *J. Geophys. Res.*, 78 (1), 227-239 (1973).
2. Sudan, R.N., J. Akinrimisi and D.T. Farley, "Generation of Small Scale Irregularities in the Equatorial Electrojet," *J. Geophys. Res.*, 78 (1), 240-248 (1973).
3. Rognlien, T.D. and J. Weinstock, "Theory of the Nonlinear Spectrum of the Gradient Drift Instability in the Equatorial Electrojet," *J. Geophys. Res.* 79 (31), 4733-4746 (1974).
4. Sudan, R.N. and M. Keskinen, "Theory of Strongly Turbulent Two-Dimensional Convection of Low Pressure Plasma," *Phys. Rev. Lett.*, 38 (17), 966-970 (1977).
5. Ferch, R.L. and R.N. Sudan, "Numerical Simulations of Type II Gradient Drift Irregularities in the Equatorial Electrojet," *J. Geophys. Res.*, 82 (16), 2283-2288 (1977).
6. McDonald, B.E., T.P. Coffey, S. Ossakow and R.N. Sudan, "Preliminary Report of Numerical Simulation of Type II Irregularities in the Equatorial Electrojet," *J. Geophys. Res.*, 79 (16), 2551-2554 (1974).
7. McDonald, B.E., T.P. Coffey, S.L. Ossakow and R.N. Sudan, "Numerical Studies of Type II Equatorial Electrojet Irregularity Development," *Radio Sci.*, 10 (3), 247-254 (1975).
8. Sato, T. and T. Ogawa, "Self-Consistent Studies of Two-Dimensional Large Scale ( $\sim 100$  m) Electrojet Irregularities," *J. Geophys. Res.*, 81 (19), 3248-3256 (1976).
9. See the preceding two articles, Barone, S., "A New Approach to Some Non-linear Plasma Turbulence Problems," and Barone, S., N. Marcuvitz, R. Pascone and N. Solimene, "Preliminary Report of Numerical Simulation of Type II Irregularities in the Equatorial Electrojet."
10. Kraichnan, R., "Irreversible Statistical Mechanics of Incompressible Hydro-magnetic Turbulence," *Phys. Rev.* 109, 1407-1422 (1958).
11. Kraichnan, R.H., "The Structure of Isotropic Turbulence at Very High Reynolds Numbers," *J. Fluid Mech.*, 5, 497-543 (1959).
12. Leslie, D.C., "Developments in the Theory of Turbulence," Oxford University Press, New York (1973).
13. Kadomtsev, B.B., "Plasma Turbulence," Academic Press, New York (1965).
14. Ott, E., and D.T. Farley, "The  $k$  Spectrum of Ionospheric Irregularities," *J. Geophys. Res.*, 79 (16), 2469-2472 (1974).
15. Prakash, S.B., H. Subbaraya and S.P. Gupta, "Investigation of the Daytime Lower Ionosphere over the Equator Using Langmuir Probe and Plasma Noise Probe," *J. Atmos. Terr. Phys.* 33 (2), 129-135 (1970).

## HIGH POWER MICROWAVE PROPAGATION THROUGH THE ATMOSPHERE

N. Marcuvitz and N. Solimene

This report summarizes the current status of our study of microwave breakdown and pulse propagation phenomena. The model used to investigate the dependence on power, density, pulse width, altitude, frequency, etc., incorporates experimental data pertinent to the ionization, recombination and energy transfer processes. The data is represented by empirical formulas. Some uncertainty is introduced by the spread of values from different sources, the transfer of data relevant to one set of conditions to other situations and the extrapolation beyond the range of measurements. The model also is restricted in that plane wave propagation in the forward scattering approximation has been used. Thus, with respect to propagation, it is only valid when the reflected power is negligible, dispersion is weak and diffraction spreading is neglected. Within the limitations of the model:

- (a) energy densities of 1-50 millijoules/cm<sup>2</sup> may be propagated to distances of 5,000 feet with a loss of less than one-half the pulse energy;
- (b) electron densities exceeding 10<sup>8</sup>/cm<sup>3</sup> may be created over the same distance (much larger densities over much shorter distances are of course also possible).

The above conclusions may be in error by significant factors because of the limitations of the model. Thus, it is exceedingly important to devise experiments to confirm the computer modeling.

The model is given by equations in a form suitable for a mixed boundary-initial value problem:

$$\frac{\partial P}{\partial z} = -\beta P \quad (1)$$

$$\frac{\partial N}{\partial t} = \alpha N - \gamma N^2 \quad (2)$$

$$\frac{\partial U}{\partial t} = \frac{\beta'}{N} P - (\alpha - \gamma N)(U + U_i) - \nu_e(U - 1) \quad (3)$$

where  $t$  is a local time, i. e., time zero propagates with the velocity of light (unity in normalized units). In this formulation  $P$ , the normalized power density of the radiation, is given at  $z=0$  for all  $t$ , while  $N$ , the electron density, and  $U$ , the normalized average electron energy, are given at  $t=0$  for all  $z \geq 0$ . That is,  $P$  is specified at a front surface (the antenna) for all time and  $N$  and  $U$  are specified initially over the entire half-space (in front of the antenna). As before, Eqs. (1) and (2) may be used with all coefficients and rates given as functions  $E/p$  and  $p$ . Alternatively, when Eq. (3) is included, the coefficients must be given as functions of  $U$  and  $p$ . The empirical values for these coefficients are given in Section A.

Some typical results for 1 cm microwave radiation are presented in the form of computer generated plots. Each plot shows the power pulse shape as a function of time at the front surface and at four equally spaced distances as solid line curves. The mid range and final range in feet are indicated along the z axis. Zero time, located at the center of the pulse at  $z = 0$ , propagates with the velocity of light. Thus the abscissa in nanosecond units is the local time at each of the distances for which results are displayed. The pulse width may be read off the abscissa. The ordinate is labeled in units appropriate for the power. Since only upper case characters can be displayed, the power  $P(z, t)$  is designated POW while the pressure  $p$  is designated P on these plots. The peak value of POW relative to breakdown power is given in the legend at the top. Also given is the pressure, P, in Torr. The dotted curve, at each distance, is the electron average energy  $U$  relative to the neutral gas average energy. This quantity has been arbitrarily scaled for display purposes. The scale may be determined by referring to the value of TMAX in the legend. This is given in eV and corresponds to the maximum value of  $U$ , for the largest distance displayed, multiplied by  $1/40$  eV, i.e., the assumed neutral gas average energy. At the largest distance, the second dotted curve represents the electron density again arbitrarily scaled. The scale is given by the value of NMAX in the legend. The units are electrons per cubic centimeter. The third line in the legend gives the peak power in  $\text{W}/\text{cm}^2$  as well as the pulse energy in  $\text{J}/\text{cm}^2$ . The pulse energy applies to the pulse at the largest displayed distance. The values for all other parameters are discussed.

Some results for  $P = 100$  Torr are given in Figs. 1(a), (b), (c). Note that the initial pulse is approximately 30 ns wide. For  $\text{POW} = 10$ , the pulse width is reduced by perhaps one third in 5,000 feet. For  $\text{POW} = 5$ , the pulse is not noticeably changed. For  $\text{POW} = 12$ , the electron density at the front surface is  $1.19 \times 10^{13}/\text{cm}^3$  which exceeds the critical value of  $1.12 \times 10^{13}/\text{cm}^3$  and therefore the pulse can not propagate within the range of validity of the model. Notice that for  $\text{POW} = 10$ , NMAX is only  $2.34 \times 10^7/\text{cm}^3$  at 5,000 feet. At the front surface the value is  $3.40 \times 10^{10}/\text{cm}^3$ . Table I gives further indication of the distance dependence. A comparison of the pulse energies at 0 and 5,000 feet confirms the estimate made above concerning the reduction in pulse width.

Figures 2(a) and 2(b) show results for a 30 ns pulse at 380 Torr for POW equal to 2 and 4 respectively. Figure 2(c) indicates that at  $\text{POW} = 5$ , the pulse cannot propagate within the range of validity of the model. When Fig. 2(b) is compared to Fig. 1(c), it may be noticed that at 380 Torr the pulse energy is about five times larger than at 100 Torr for nearly similar attenuation, i.e., similar electron densities. This is qualitatively accounted for by the fact that the breakdown power increases as the square of the pressure while the collisional frequencies increase linearly with pressure.

## WAVE-MATTER INTERACTIONS

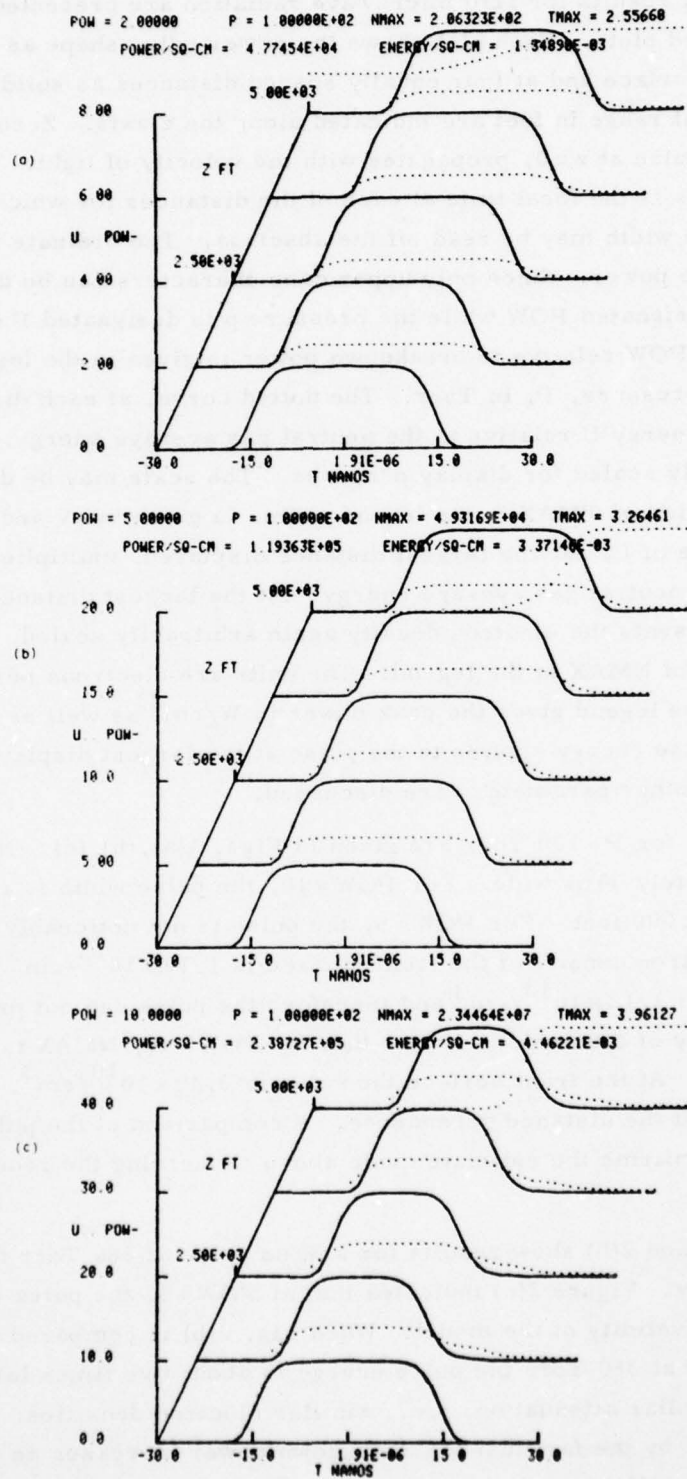


Fig. 1. Propagation through atmosphere.

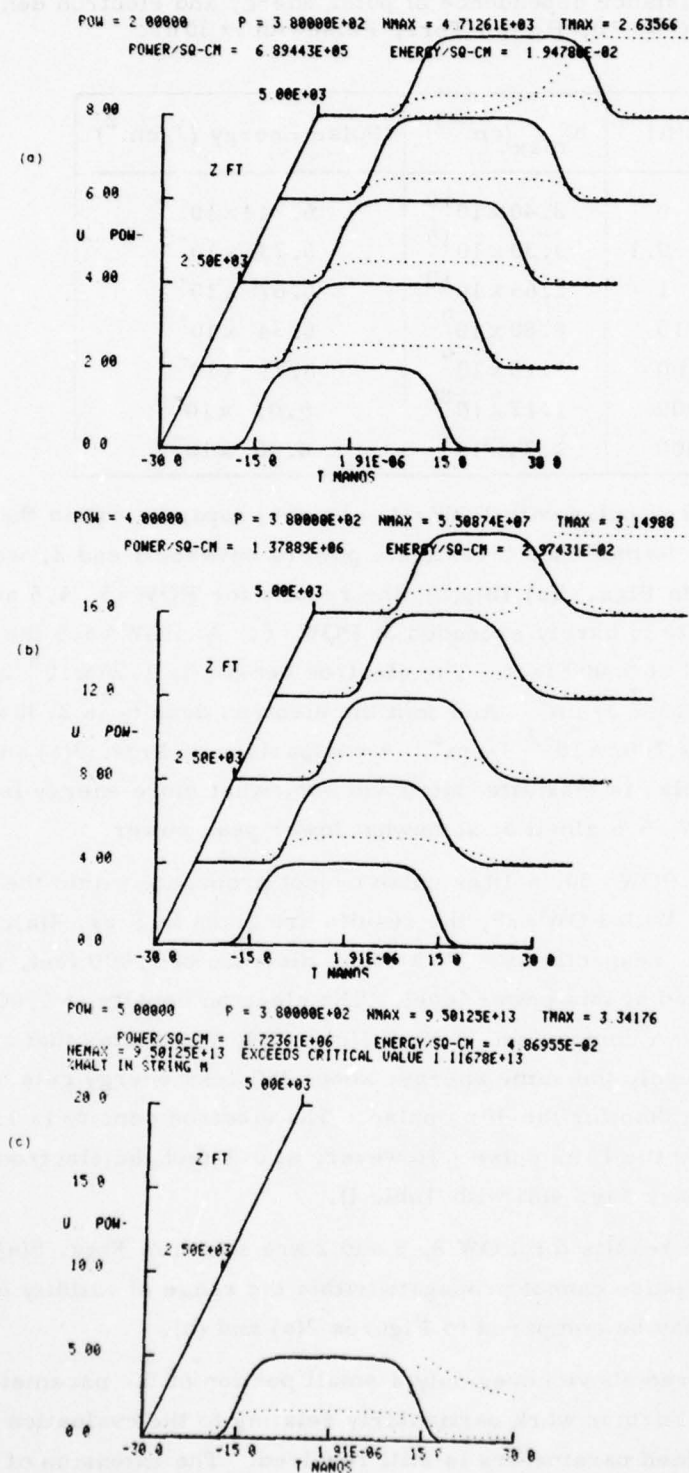


Fig. 2. Propagation through atmosphere.

TABLE I. Distance dependence of pulse energy and electron density  
 POW = 10, P = 100 Torr, Pulsewidth = 30 ns.

z(ft)	$N_{\max}(\text{cm}^{-3})$	Pulse Energy ( $\text{J}/\text{cm}^2$ )
0	$3.40 \times 10^{10}$	$6.744 \times 10^{-3}$
0.1	$3.30 \times 10^{10}$	$6.736 \times 10^{-3}$
1	$2.65 \times 10^{10}$	$6.67 \times 10^{-3}$
10	$8.80 \times 10^9$	$6.34 \times 10^{-3}$
100	$1.13 \times 10^9$	$5.71 \times 10^{-3}$
1000	$1.17 \times 10^8$	$5.03 \times 10^{-3}$
5000	$2.34 \times 10^7$	$4.46 \times 10^{-3}$

At 760 Torr a 30 ns pulse with POW = 2, can not propagate within the range of validity of the model. Rather than investigate powers between 1 and 2, we changed the pulse width to 10 ns. In Figs. 3(a), (b), (c), the results for POW = 5, 5.5 and 6 are shown. Note the critical density is barely exceeded at POW = 6. At POW = 5.5 the pulse is significantly shortened at 5,000 feet. The electron density is  $1.29 \times 10^8/\text{cm}^3$  and the pulse energy is  $4.81 \times 10^{-2} \text{ J}/\text{cm}^2$ . At 1 foot the electron density is  $2.32 \times 10^{11}/\text{cm}^3$  and the pulse energy is  $7.02 \times 10^{-2} \text{ J}/\text{cm}^2$ . A comparison of Figs. 3(a) and 3(b) indicates that at POW = 5, the pulse is less attenuated and somewhat more energy is available at 5,000 feet than at POW = 5.5 albeit at somewhat lower peak power.

At 100 Torr and POW = 30, a 10 ns pulse cannot propagate within the range of validity of the model. With POW = 29, the results are given in Figs. 4(a), (b), (c) for 0.1, 10 and 5,000 feet, respectively. To attain a distance of 5,000 feet, about half the pulse must be sacrificed at this power level. The electron density at 5,000 is again of the order of  $10^7/\text{cm}^3$ . A comparison of Fig. 1(c) and 4(c) indicates that although both pulses started with roughly the same energy, about 25% less energy gets out to 5,000 feet for the 10 ns pulse than for the 30 ns pulse. The electron density is 1.5 times larger at 5,000 feet for the 10 ns pulse. However, at 0.1 feet the electron density is 50 times larger (compare Fig. 4(a) with Table I).

At 380 Torr, the results for POW 8, 5 and 2 are shown in Figs. 5(a), (b) and (c). For POW = 10, a 10 ns pulse cannot propagate within the range of validity of the model at 380 Torr. These may be compared to Figures 2(a) and (b).

The plots discussed above cover only a small portion of the parameter range which may be investigated. Further work particularly relating to the evaluation of the validity of empirically determined parameters is still required. The extension of the model



## WAVE-MATTER INTERACTIONS

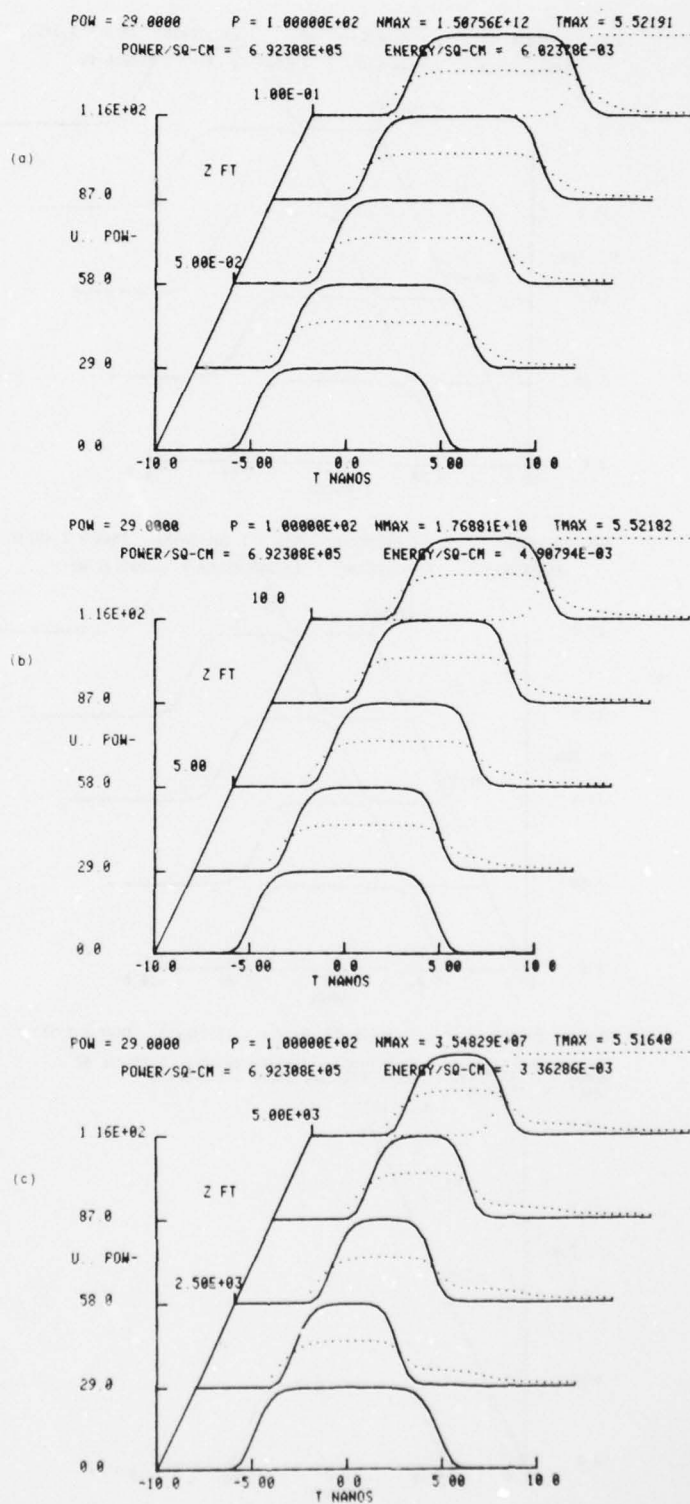


Fig. 4. Propagation through atmosphere.

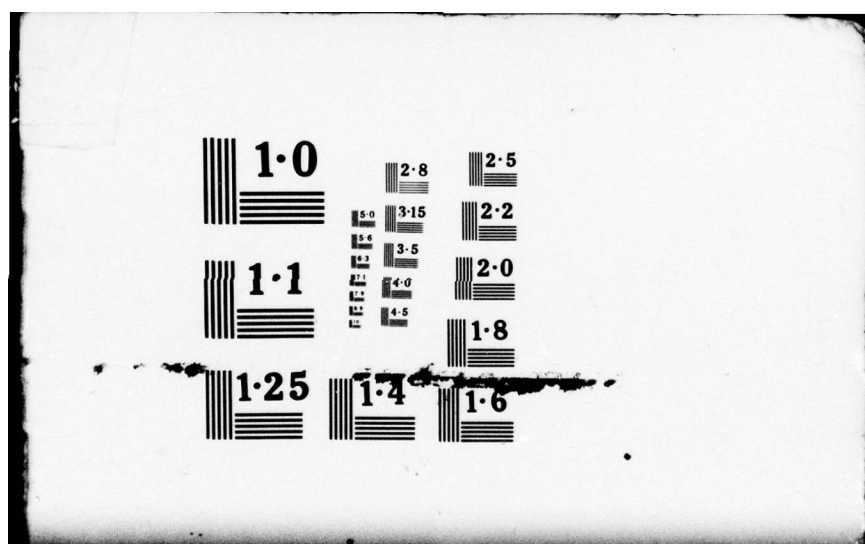
AD-A063 181

POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 9/3  
PROGRESS REPORT NUMBER 43 TO THE JOINT SERVICES TECHNICAL ADVIS--ETC(U)  
NOV 78 A A OLINER F44620-78-C-0074  
POLY-MRI-452.43-78 NL

UNCLASSIFIED

4 OF 6  
AD A  
063181





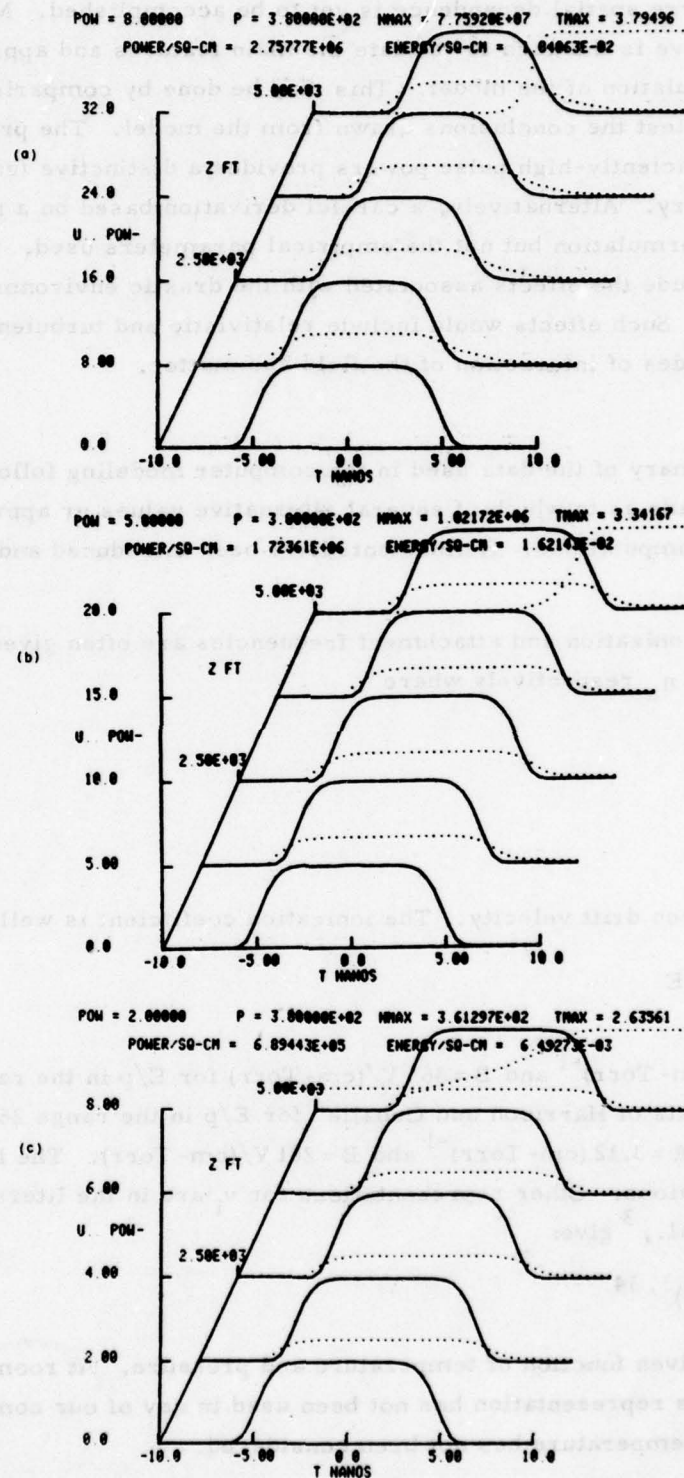


Fig. 5. Propagation through atmosphere.

to include transverse spatial dependence is yet to be accomplished. More important than any of the above is the need to validate the main features and approximations involved in the formulation of the model. This may be done by comparison with experiments designed to test the conclusions drawn from the model. The predicted pulse shortening for sufficiently-high pulse powers provides a distinctive feature to test the validity of the theory. Alternatively, a careful derivation based on a microscopic theory may validate the formulation but not the empirical parameters used. Such a treatment must properly include the effects associated with the drastic environment provided by the intense fields. Such effects would include relativistic and turbulent modifications of the ordinary modes of interaction of the field and matter.

#### A. Appendix

A brief summary of the data used in the computer modeling follows. A number of choices were made as to which of several alternative values or approximations was to be used in the computations. Refinements have been introduced and are continuing to be introduced.

The data on ionization and attachment frequencies are often given in terms of the coefficients  $\eta_i$  and  $\eta_a$  respectively where

$$\nu_i = v_d \eta_i$$

and

$$\nu_a = v_d \eta_a$$

and  $v_d$  is the electron drift velocity. The ionization coefficient is well represented by

$$\frac{\eta_i}{p} = A e^{-Bp/E}$$

where for  $A = 15 \text{ (cm-Torr)}^{-1}$  and  $B = 365 \text{ V/(cm-Torr)}$  for  $E/p$  in the range 100-800 V/(cm-Torr).<sup>1</sup> The data of Harrison and Geballe<sup>2</sup> for  $E/p$  in the range 25-65 V/(cm-Torr) may be fitted with  $A = 3.12 \text{ (cm-Torr)}^{-1}$  and  $B = 201 \text{ V/(cm-Torr)}$ . The latter has been used in our calculations. Other representations for  $\nu_i$  are in the literature. For example, Mayhan et al.,<sup>3</sup> give:

$$\nu_i = f(T, p) \left( \frac{E}{p} \right)^{5.34}$$

where  $f(T, p)$  is a given function of temperature and pressure. At room temperature  $f = 8.35 \times 10^{-4}$ . This representation has not been used in any of our computations, since dependence on air temperature has not been considered.

The attachment coefficient data is somewhat less certain. The data of Harrison and Geballe<sup>2</sup> for  $E/p$  in the range 25-60 V/(cm-Torr) and that of Chatterton and Craggs<sup>4</sup> in the range 2-30 V/(cm-Torr) are in fair agreement over the overlapping region  $25 < E/p < 30$ . For the first calculations, a rough compromise was struck by using

$$\frac{\eta_a}{p} = a + b \left(\frac{E}{p}\right)^2$$

with  $a = 5.68 \times 10^{-4} (\text{cm-Torr})^{-1}$  and  $b = 3.05 \times 10^{-6} (\text{cm-Torr/V})^2 / (\text{cm-Torr})$ . This was not entirely satisfactory since when extrapolated to sufficiently high values of  $E/p$ ,  $\eta_a/p$  exceeded the corresponding values of  $\eta_i/p$ . For the later calculations we allowed  $\eta_a/p$  to saturate. This is physically plausible and avoids the difficulty mentioned above. In particular we used

$$\frac{\eta_a}{p} = a - b e^{-(E/15p)^2}$$

with  $a = 4.0 \times 10^{-3} (\text{cm-Torr})^{-1}$  and  $b = 3.5 \times 10^{-3} (\text{cm-Torr})^{-1}$ . In any event, attachment should not be very important when  $E/p$  exceeds the breakdown value by more than a factor of two or so.

The drift velocity data used is that of Emeleus, Lunt and Meek<sup>5</sup> for  $E/p$  in the range 20-595 V/(cm-Torr). This data is well summarized by:

$$v_d = 0.125 \times 10^7 \left(\frac{E}{p}\right)^{2/3} \text{ cm/sec}$$

where  $E/p$  is given in V/(cm-Torr). Since

$$v_d = \frac{e}{mv} E,$$

we may infer the momentum transfer collision frequency

$$\frac{\nu}{p} = 1.41 \times 10^9 \left(\frac{E}{p}\right)^{1/3} (\text{sec-Torr})^{-1}$$

A cruder fit to this data is:

$$v_d = a_0 + a_1 \left(\frac{E}{p}\right)$$

where  $a_0 = 1.44 \times 10^7 \text{ cm/sec}$  and  $a_1 = 1.15 \times 10^5 \text{ cm}^2 \cdot \text{Torr} / (\text{sec} \cdot \text{V})$ . For simplicity, the latter was used in the first calculations, and the momentum transfer collision frequency was taken to be constant and equal to  $5 \times 10^9 (\text{sec-Torr})^{-1}$ . For the later calculations, the more accurate fit was used for  $v_d$  and the corresponding momentum transfer collision frequency was used. Other data and/or derived values are in the literature. For

example, Felsenthal<sup>6</sup> reports

$$\frac{\gamma}{p} = 5.3 \times 10^9 + 1.115 \times 10^7 \left( \frac{E}{p} \right) (\text{sec-Torr})^{-1}$$

for  $E/p$  in  $V/(\text{cm-Torr})$ . The author does not however consider this to be a very accurate value. Also Felsenthal and Proud<sup>7</sup> use

$$v_d = 7 \times 10^6 + 2 \times 10^6 \left( \frac{E}{p} \right) \text{cm/sec}$$

which is adopted from Nielson and Bradbury.<sup>8</sup> Ryzko<sup>9</sup> reports

$$v_d = 5 \times 10^6 + 256 \times 10^3 \left( \frac{E}{p} \right) \text{cm/sec}$$

for the range  $E/p = 50$  to  $100 V/(\text{cm-Torr})$ .

The recombination coefficient was taken from the data of Sayers as given in S. C. Brown's book.<sup>10</sup> The value of  $\gamma$  is approximately  $2.3 \times 10^{-6} (\text{cm}^3/\text{sec})$  at one atmosphere and decreases to zero as the pressure is lowered. We therefore approximated:

$$\frac{\gamma}{p} \approx \frac{2.3 \times 10^{-6}}{760} \text{cm}^3/(\text{sec-Torr})$$

In order to convert the empirical expressions given above to functions of the electron average energy it is necessary to relate  $E/p$  to  $U$ . This can be done for steady-state conditions below breakdown. It is then necessary to extrapolate the result to transient conditions well above breakdown. When this is done using the data of Huxley and Zazou<sup>11</sup> we obtain

$$U = 18.7 \left( \frac{E}{p} \right)^{2/3}$$

for  $E/p < 25 V/(\text{cm-Torr})$ . The data of Crompton et al., yields<sup>12</sup>

$$U = 15 \left( \frac{E}{p} \right)^{1/2}$$

for  $E/p < 20 V/(\text{cm-Torr})$ . Each of these expressions represents the experimental data to within 20%. We elected to use the latter expression.

The other parameter which must be determined is  $\nu_e$ , the energy transfer collisional frequency. The first calculations with this parameter used the value of  $2m/M \approx 10^{-4}$ , i. e., the ratio of electron mass and neutral gas mass. This should adequately model the transfer of kinetic energy. It was soon found to be a poor approximation and led to excessively high values of  $U$  with the consequence that extremely high ionization rates were introduced in the course of the computation. The cure for this is found when it is realized that the air molecules can absorb energy from the electron

in the form of rotational and vibrational excitation as well as kinetically. Such processes are dependent on the electron energy and therefore  $v_e$  may be expected to depend on  $U$ . The data in Ref. 12 shows a dependence for  $v_e$  on  $E/p$  which is roughly given by

$$\frac{v_e}{v} = 1.17 \times 10^{-3} \left( \frac{E}{p} \right)$$

for  $2 \leq \frac{E}{p} \leq 20$  V/(cm-Torr). If we extrapolate this to much larger values of  $E/p$  it must obviously fail when the ratio  $v/v_e$  exceeds unity. To avoid this, we have assumed that  $v_e/v$  saturates for large  $E/p$  to a value of  $1 \times 10^{-1}$ . With this rather arbitrary choice, we have elected to use

$$\frac{v_e}{v} = 1 \times 10^{-1} [1 - e^{-1.17 \times 10^{-2} (E/p)}]$$

in our later calculations. A limited check for values of  $E/p$  below 30 (the breakdown value) is provided by calculating  $U$  and comparing to the values given in Reference 12. This comparison is made in Table II, which also includes the value obtained from the

TABLE II. Comparison of average energy calculations.

$E/p$ V/(cm-Torr)	$U_{\text{CHS}}$ Ref. 16	$U_{\text{HZ}}$ Ref. 15	$U_{\text{CALC}}$	$15 \left( \frac{E}{p} \right)^{1/2}$
5	40.3		37.2	33.2
10	47.7	86.0	51.2	47.2
15	53.0	108.5	61.6	58.0
20	59.5	128.0	70.4	67.2

expression relating  $U$  to  $E/p$ . The agreement is only fair between  $U_{\text{CHS}}$  and  $U_{\text{CALC}}$ . More significant is the consistency of  $U_{\text{CALC}}$  and  $15 (E/p)^{1/2}$ . These values were in drastic disagreement when  $v_e/v$  was approximated as  $2m/M$ . Also note the discrepancy between the results of References 11 and 12.

Naval Surface Weapons Center  
N60921-77-C-A303

N. Solimene

#### REFERENCES

1. A. von Engel, Encyclopedia of Physics, Vol. 21, Springer (1956).
2. M.A. Harrison and R. Geballe, Phys. Rev. 91, 1 (1953).
3. J. T. Mayhan et al., J. Appl. Phys. 42, 5362 (1971).
4. P.A. Chatterton and J.D. Craggs, Proc. Phys. Soc. 85, 355 (1965).
5. K.G. Emeleus, R.W. Lunt and C.A. Meek, Proc. Roy. Soc. A, 156, 394 (1936).

6. P. Felsenthal, J. Appl. Phys. 37, 4557 (1966).
7. P. Felsenthal and J.M. Proud, Phys. Rev. 139, A1796 (1965).
8. R.A. Nielson and N.E. Bradbury, Phys. Rev. 51, 69 (1937).
9. H. Ryzko, Proc. Phys. Soc. 85, 1283 (1965).
10. S.C. Brown, "Basic Data of Plasma Physics," 2nd ed., MIT Press (1966).
11. L.H.G. Huxley and A.A. Zaazon, Proc. Roy. Soc. A, 196, 402 (1949).
12. R.W. Crompton, L.G. Huxley and D.J. Sutton, Proc. Roy. Soc. A, 218, 507 (1953).

TABLE II  
Comparison of average energy calculations

$E_p V / (kT_e - T_e)$ Ref. 12	$U_{CHS}$ Ref. 12	$U_{CHS}$ Ref. 12	$U_{CHS}$ Ref. 12	$U_{CHS}$ Ref. 12
5	4.1	4.1	4.1	4.1
10	47.7	40.8	47.7	47.7
15	130.0	98.5	130.0	130.0
20	197.0	118.0	197.0	197.0

expression relating  $U$  to  $E_p$ . The agreement is only fair between  $U_{CHS}$  and  $U_{CALC}$ . More significant is the consistency of  $U_{CHS}$  and  $U_{CALC}$ . These values were 10% different in the case of  $U_{CHS}$ . Also note the discrepancy between the results of References 11 and 12.

Naval Surface Warfare Center  
Navy-77-C-4311

#### REFERENCES

1. A. von Engel, Encyclopedia of Physics, Vol. 31, Springer (1958).
2. M.A. Harrison and R. Gosselin, Phys. Rev. 171, 1199 (1968).
3. J.T. Mather et al., J. Appl. Phys. 45, 3558 (1974).
4. P.A. Christen and J.D. Cragg, Proc. Phys. Soc. 85, 1283 (1965).
5. R.G. Fennel, R.W. Lant and C.A. Meek, Phys. Rev. 179, 119 (1959).

## KINETIC THEORY TREATMENT OF METAL EVAPORATION FRONT

M. Newstein, N. Solimene and J. Hammer

We are concerned with the evaporation of metal vapor from a surface exposed to laser radiation. It is assumed that the vapor density is small enough that the laser energy is absorbed at the metal surface rather than in the vapor. There is a transition layer extending several mean free paths from the metal surface in which conditions deviate significantly from local equilibrium. The rate of evaporation depends on the hydrodynamic parameters, density, mean velocity and temperature, beyond this transition layer. The relationship between these parameters and the temperature of the metal surface has been discussed according to various models.<sup>1,2</sup>

Anisimov<sup>2</sup> assumed a form for the Boltzmann distribution function in the region of the evaporation front which contained features of the physics of the evaporation process. Using this he was able to relate the conditions at the surface of the metal to those on a plane outside the transition layer. He concluded that the flow of atoms condensed back on the surface of the metal amounts to 18% of the flow of the vaporized atoms.

In this paper we study the structure of the evaporation front using a method which has been useful for determining shock wave structure.<sup>3</sup> Less restrictive assumptions are made than those of Anisimov, hence we are required to solve the Boltzmann equation for the distribution function which evolves from the surface distribution. In the relaxation time approximation,<sup>4</sup> and under the assumption of no spatial variation transverse to the x-axis, the Boltzmann equation is written:

$$\frac{\partial f}{\partial t} + v_x \frac{\partial f}{\partial x} = \frac{1}{\tau} (\psi - f) \quad (1)$$

where  $\psi/\tau$  and  $f/\tau$  are the collisional gain and loss terms. We assume that the collision time  $\tau$  is a function of the moments of the distribution function, and that the function  $\psi$  toward which the distribution function relaxes has the form:

$$\psi(x, \vec{v}, t) = n(x, t) \left[ \frac{m}{2\pi kT(x, t)} \right]^{3/2} e^{-m \frac{[\vec{v} - u(x, t)]^2}{2kT(x, t)}} \quad (2)$$

where the density,  $n$ , mean velocity,  $u$ , and temperature,  $T$ , are determined by the appropriate velocity integrals of the distribution function,  $f$ :

$$n(x, t) = \iiint d^3v f(x, \vec{v}, t) \quad (3a)$$

$$n(x, t) u(x, t) = \iiint d^3v v_x f(x, \vec{v}, t) \quad (3b)$$

$$3n(x, t) \frac{kT(x, t)}{m} = \iiint d^3v \{ [v_x - u]^2 + v_y^2 + v_z^2 \} f(x, \vec{v}, t) \quad (3c)$$

The form of Eq. (2) together with Eq. (3) guarantees that the relaxation term of Eq. (1) describes two body collisions which conserve energy, momentum and particle number. The spatial domain of interest extends from the metal surface,  $x=0$ , at which the time-varying temperature is  $T_0(t)$ , to a plane  $x=l$ , beyond which conditions of local equilibrium obtain. The corresponding boundary conditions are:

$$f(x=0, \vec{v}, t) = n_0(t) \left[ \frac{m}{2\pi kT_0(t)} \right]^3 e^{-\frac{mv^2}{2kT_0(t)}} + R f(x=0, -v_x, v_y, v_z) \quad \text{for } v_x > 0 \quad (4a)$$

and

$$f(x=l, \vec{v}, t) = \psi(l, \vec{v}, t) \quad \text{for } v_x < 0 \quad (4b)$$

Equation (1), subject to the boundary conditions Eq. (4) and the initial condition  $f(t=0) = 0$ , may be integrated and the solution expressed as

$$f(x, \vec{v}, t) = f(x=0, v, \text{tr}(x)) e^{-\int_0^x \frac{dx'}{v_x \tau(x', \text{tr}(x-x'))}} + \int_0^x \frac{dx'}{v_x \tau(x', \text{tr}(x-x'))} \psi(x', \vec{v}, \text{tr}(x-x')) e^{-\int_{x'}^x \frac{dx''}{v_x \tau(x'', \text{tr}(x-x''))}} \quad \text{for } v_x > 0 \quad (5a)$$

and

$$f(x, \vec{v}, t) = f(x=l, \vec{v}, \text{tr}(x)) e^{-\int_x^l \frac{dx'}{v_x \tau(x', \text{tr}(x'-x))}} + \int_x^l \frac{dx'}{v_x \tau(x', \text{tr}(x'-x))} \psi(x', \vec{v}, \text{tr}(x'-x)) e^{-\int_{x'}^l \frac{dx''}{v_x \tau(x'', \text{tr}(x''-x))}} \quad \text{for } v_x < 0 \quad (5b)$$

where

$$\text{tr}(\mathbf{x}) = t - \left| \frac{\mathbf{x}}{v_{\mathbf{x}}} \right| \quad (5c)$$

and where the function  $f$  and  $\psi$  on the right hand side of Eq. (5) vanish when their third argument becomes negative. The density  $n_0(t)$  and surface temperature  $T_0(t)$  are determined by the net energy flux across the metal surface and the equilibrium vapor pressure. Thus, assuming that all the absorption occurs at the metal surface, we have:

$$T_0(t) = T_0(t_0) + \frac{\sqrt{c/\pi}}{K} \int_{t_0}^t \mathcal{J}(t - \tau) \frac{d\tau}{\sqrt{\tau}} \quad (6)$$

where  $c$  and  $K$  are the thermal diffusivity and conductivity of the metal and  $\mathcal{J}$  is the net energy flux across the surface given by:

$$\mathcal{J} = g(t) - \left[ \frac{m}{2} \int d^3 v v_{\mathbf{x}} \left[ v_{\mathbf{x}}^2 + v_{\mathbf{y}}^2 + v_{\mathbf{z}}^2 \right] f(\mathbf{x} = 0, \vec{v}, t) + n u V \right] \quad (7)$$

In Equation (7),  $g(t)$  is the absorbed laser energy flux and  $V$  is the energy of vaporization/per atom. The equilibrium density  $n_0$  can be expressed in terms of the temperature by an expression of the form

$$n_0 = \frac{A}{T} e^{-\frac{V}{kT}}, \quad (8)$$

where  $A$  is an empirically determined constant.

The evaluation of these coupled equations for realistic situations required the use of numerical methods. The interests of speed and flexibility would be served if the dependence on the transverse velocity coordinates could be integrated out analytically. This cannot be done exactly, and the results of one approximate procedure will be described here. Integrating Eq. (1) over the transverse coordinate  $v_{\mathbf{y}}$  and  $v_{\mathbf{z}}$  and defining the reduced distribution functions by:

$$F(\mathbf{x}, v_{\mathbf{x}}, t) = \iint dv_{\mathbf{y}} dv_{\mathbf{z}} f(\mathbf{x}, \vec{v}, t) \quad (9a)$$

$$\Psi(\mathbf{x}, v_{\mathbf{x}}, t) = \iint dv_{\mathbf{y}} dv_{\mathbf{z}} \psi = n \left( \frac{m}{2\pi kT} \right)^{1/2} e^{-\frac{m(v_{\mathbf{x}} - u)^2}{2kT}} \quad (9b)$$

we get

$$\frac{\partial F}{\partial t} + v_x \frac{\partial F}{\partial x} = \frac{1}{\tau} [\Psi - F] \quad (10)$$

The temperature  $T$ , which appears in  $\Psi$ , cannot be found from the reduced distribution function  $F$ , rather we have

$$3T = T_\ell + 2T_t, \quad (11)$$

where

$$\begin{aligned} n \frac{kT_\ell}{m} &= \int dv_x (v_x - u)^2 F \\ n \frac{kT_t}{m} &= \iiint d^3v v_y^2 f = \iiint d^3v v_z^2 f \end{aligned} \quad (12)$$

Thus, to close our reduced equations, we require another equation (and boundary conditions) for the transverse temperature  $T_t$ . This may be found from Eq. (1) by multiplying  $v_y^2$  and integrating. A further approximation

$$\iiint v_x v_y^2 f d^3v \approx nu \frac{kT_t}{m}, \quad (13)$$

(i.e. the neglect of the longitudinal conductivity of heat associated with the transverse thermal motion) yields an approximate equation for  $T_t$ :

$$\frac{\partial}{\partial t} n T_t + \frac{\partial}{\partial x} nu T_t = \frac{n}{3\tau} (T_\ell - T_t) \quad (14)$$

The boundary condition used is  $T_t(0) = T_o$ . The results for a calculation run to the steady state are shown in Figure (1) and (2). Figure (1) illustrates the dependence of the steady state reduced distribution function on  $v_x$  and  $x$ . At the input plane  $x=0$ , the function for positive  $v_x$  has the form of a Maxwellian at the surface temperature. For negative values of  $v_x$  the surface distribution function is determined by the molecules scattered from planes with  $x > 0$ . As the coordinate  $x$  increases from  $x=0$ , the distribution function tends toward the local equilibrium form of a Gaussian centered at the mean velocity  $u(x)$ . Figure 2 illustrates the steady state dependence of the moments  $n$ ,  $u$ ,  $nu$ ,  $c_e$  and  $c_t$  on  $x$ , where

$$c_e = \sqrt{\frac{5}{3} \frac{kT_e}{m}}; \quad c_t = \sqrt{\frac{5}{3} \frac{kT_t}{m}} \quad (15)$$

can be identified with the longitudinal and transverse sound velocities associated with the corresponding temperatures. The clearest indication of the deviation from conditions

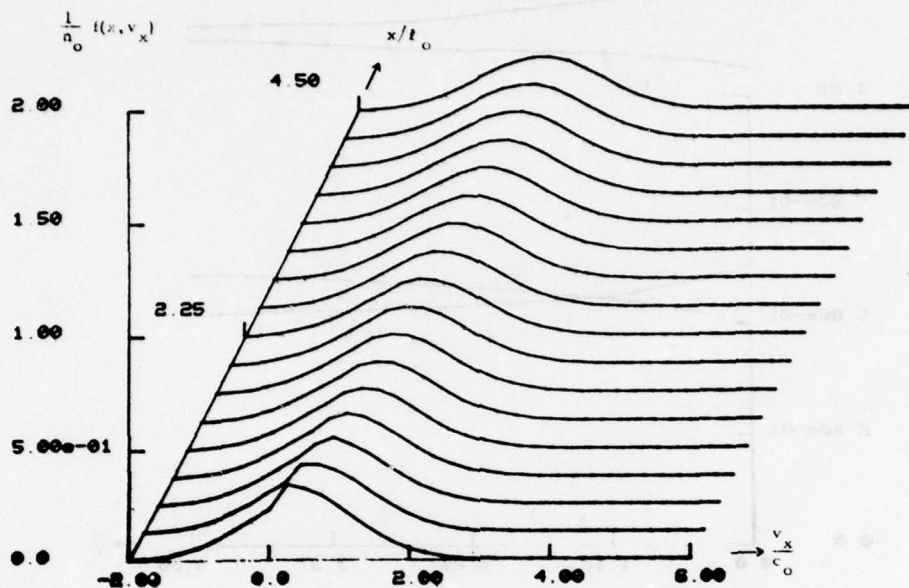


Fig. 1. This figure illustrates the steady state distribution function  $f(x, v_x)$  as a function of velocity  $v_x$ , and distance from the metal surface,  $x$ . The temperature of the metal surface  $T_0$  would, under equilibrium conditions, lead to a number density  $n_0$  and sound velocity  $C_0 = \sqrt{kT_0/m}$ . These quantities are used to normalize the variable  $f$  and  $v_x$ . The collision frequency is taken to be given by  $1/\tau = nc\sigma$  and the distance variable  $x$  is normalized relative to the equilibrium mean free path  $\ell_0 = 1/n_0\sigma$ . At the surface  $x=0$  the distribution function is Maxwellian for positive values of  $v_x$  for negative values it assumes a form determined by the effect of collisions to the right of the  $x=0$  plane. As  $x$  increases, the distribution function assumes the local equilibrium form.

of local thermal equilibrium is provided by the difference between  $c_e$  and  $c_t$ . Beyond a plane about 4.5 mean free paths from the surface the values of  $c_e$ ,  $c_t$  come together. This is the plane beyond which a hydrodynamic description becomes valid.

National Science Foundation  
ENG76-21829A01

M. Newstein, N. Solimene and J. Hammer

## WAVE-MATTER INTERACTIONS

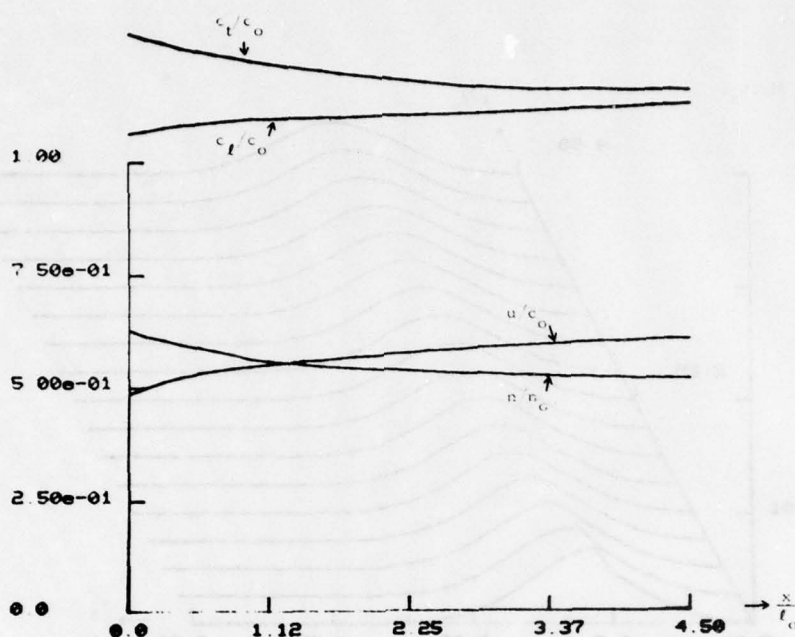


Fig. 2. Illustrates the kinetic equation solutions for the steady-state spatial dependence of longitudinal and transverse temperature  $T_l/T_0 = (c_l/c_0)^2$ ,  $T_t/T_0 = (c_t/c_0)^2$ , mean velocity  $u$ , and number density  $n$ . Our model of collisions causes relaxation toward one temperature  $T = 2/3 T_c + 1/3 T_l$ . In the region where  $c_t/c_e$  condition of local equilibrium do not exist and the hydrodynamic equations are inapplicable. For values of  $x$  greater than several mean free paths the hydrodynamic equations may be used to obtain the relevant moments of the distribution function with boundary conditions determined by the kinetic equation solutions.

## REFERENCES

1. Oleg N. Krötkin, "Generation of High-Temperature Vapors and Plasmas by Laser Radiation," *Laser Handbook*, edited by F. T. Arocchi and F. A. Schulz-DuBois, North-Holland Publ. Co., 1371 (1972).
2. S. I. Anisimov, "Vaporization of a Metal Absorbing Laser Radiation," *Soviet Phys. JETP* 27 182 (1968).
3. B. M. Segal and J. H. Ferziger "Shock Wave Structure by Several New Modeled Boltzmann Equations" *Proceedings of 2nd Intl. Conf. on Numerical Methods in Fluid Dynamics*, Berkeley, Calif., 279 (September 1970).
4. P. L. Bhatnagar, G. P. Gross and M. Kroop, "A Model for Collision Processes in Gases," *Phys. Rev.*, 94, 511 (1954).

## FUNCTIONAL EQUATION APPROACH TO THE THREE-WAVE NONLINEAR INTERACTION

S. T. Peng and E. S. Cassedy

The resonant (parametric) three-wave interaction has been studied in the nonlinear regime as either the initial-value problem<sup>1,2</sup> or the boundary-value problem.<sup>1-3</sup> In the initial-value problem<sup>1,2</sup> pulse envelopes of propagating waves overlap in an unbounded space for finite intervals of time either due to contra-directional propagation or due to overtaking (co-directional waves of different group velocities). The boundary-value problem,<sup>1-3</sup> on the other hand, deals with the interaction of three-waves in a bounded region of interaction, with one wave (the pump) specified to have a fixed value at one boundary attributed to a wave, of constant amplitude, incident from the external region. It is for the initial-value problem that we shall show the application of the integral equation approach.

The three-wave problem is one of several nonlinear wave problems of interest in laser optics or laser plasma interactions. The other problems include the solution of nonlinear wave equations<sup>5,6</sup> (e.g., the KdV equation or the nonlinear Schrödinger equation), several of which display soliton characteristics. Thus far in the literature on nonlinear waves the predominant analytical approach to solutions has been through the Inverse Scattering Transform<sup>5,6</sup> (IST), and numerical solutions have typically been obtained through differential techniques.<sup>1</sup> The Inverse Scattering Transform approach is a powerful method but requires an auxiliary formalism and multiple steps to achieve the wave solution and differential techniques often encounter problems of numerical stability. We have therefore explored the nonlinear integral equation as an alternative approach, both analytical and numerical.

We have considered the system of first-order nonlinear partial differential equations for the three-wave interaction:

$$\left(\frac{\partial}{\partial t} + \sigma_i + v_i \frac{\partial}{\partial x}\right) a_i = \beta_i a_j^* a_k^* \quad (1)$$

for

$$i, j, k = 1, 2, 3 \quad \text{and} \quad i \neq j \neq k$$

where  $a_i$  is the complex wave amplitude,  $\beta_i$  the nonlinear coupling coefficient,  $\sigma_i$  the damping constant and  $v_i$  the wave velocity of the  $i$ -th mode. Such a system of coupled equations has been extensively analyzed in the literature by various approximation techniques<sup>7</sup> or for special cases.<sup>2,7</sup> Here we show the approach to this system of nonlinear differential equations by the method of integral equations.

We observe that the system of nonlinear differential equations in Eq. (1) are

coupled through the nonlinear term in each equation. If the nonlinear term is viewed as a known source of excitation, the three equations in Eq. (1) then appear linear and independent of one another, and the solutions can therefore readily be written as:

$$a_i(x, t) = e^{-\sigma_i t} f_i(x - v_i t) + \beta_i \int_0^t B_{jk}(x - v_i t + v_i \tau, \tau) e^{-\sigma_i(t-\tau)} d\tau \quad (2)$$

$$B_{jk}(x, t) = a_j^*(x, t) a_k^*(x, t) \quad (3)$$

where  $f_i$  is a given function that characterizes the initial condition of the  $i$ -th mode. However, with the source functions  $B_{jk}$  given by Eq. (3), Eq. (2) actually forms a system of nonlinear integral equations from which the wave amplitudes  $a_i$  still remain to be determined. In comparison with the original differential equations, the integral equations have the advantage that the initial conditions appear explicitly in the equations; therefore, the integral equations are particularly convenient for the study of evolutions of wave packets from a given set of initial conditions.

The equations in Eq. (2) contain two terms which may be interpreted as follows: The first term is the propagation of the initial wave profile in the absence of the nonlinearity in the system. The second term accounts for the nonlinear coupling effect. Furthermore, the integral characterizing nonlinear interactions means physically that the field of the  $i$ -th mode at time  $t$  depends on the past history of the other two modes. In the presence of damping, the effect of the past history is damped by a factor  $e^{-\sigma_i(t-\tau)}$  for the  $i$ -th mode for  $0 < \tau < t$ . Evidently, in the special case of heavily damped  $i$ -th mode<sup>2</sup>,  $\sigma_i \gg 1$ , the first term in Eq. (2) is exponentially small and the integrand in the second term may have significant contribution only in the vicinity:  $\tau \approx t$ . Therefore, Eq. (2) may be approximated by:

$$a_i(x, t) \approx \beta_i a_j^*(x, t) a_k^*(x, t) \sigma_i \quad (4)$$

which means physically that the  $i$ -th mode depends on the instantaneous interaction between the other two modes at every point ( $x$ ) in space. The same result was obtained by neglecting the derivative terms<sup>2</sup> in the differential equations.

Next, in order to illustrate the integral equation formulation, we consider the case where  $a_1$  is the heavily damped mode (i.e.,  $\sigma_1 \gg 1$ ). Substituting from Eq. (4) for  $a_1$ , we obtain for  $i = 2$  and  $3$  the following:

$$\left( \frac{\partial}{\partial t} - \sigma_i + v_i \frac{\partial}{\partial x} \right) a_i = \left( \frac{\beta_1 \beta_i}{\sigma_1} \right) |a_j|^2 a_i \quad (5)$$

for  $i, j = 2$  and  $3$ ;  $i \neq j$ .

Multiplying by  $a_i^*$ , Eq. (5) becomes:

$$\left(\frac{\partial}{\partial t} - 2\sigma_i + v_i \frac{\partial}{\partial x}\right) I_i = 2\rho_i I_i I_j \quad (6)$$

$$I_i = |a_i|^2 \quad \text{and} \quad \rho_i = \beta_1 \beta_i / \sigma_1 \quad (7)$$

where  $I_i(x, t)$  is the intensity of the  $i$ -th mode and is a non-negative real function of  $x$  and  $t$ . Similarly, the differential equation, Eq. (6), may be converted into an integral equation and the result in this case is:

$$I_i(x, t) = e^{-2\sigma_i t} e^{2\rho_i \int_0^t I_j(x - v_i \tau + v_i \tau, \tau) d\tau} g_i(x - v_i t) \quad (8)$$

for  $i, j = 2$  and  $3$ , and  $i \neq j$ ,

where  $g_i$  characterizes the initial intensity of the  $i$ -th mode. Since  $I_i$  is non-negative, the integral in the exponent of Eq. (8) is a non-decreasing function of  $t$ . Therefore,  $I_i$  will be exponentially decreasing in time, if  $\sigma_i \neq 0$  and  $\beta_i > 0$ . On the other hand, if  $\sigma_i \neq 0$  and  $\beta_i > 0$ ,  $I_i$  may be increasing or decreasing in time, depending on the initial conditions.

Equations (2) and (8) are nonlinear integral equations from which the amplitude (or intensity) of the  $i$ -th mode is to be determined. Although such integral equations have been shown here to be useful for the study of general nonlinear interactions among the three waves, they are, however, not amenable to direct exact solutions. For a numerical analysis of the nonlinear wave evolution, it is desirable to approximate the nonlinear integral equation, for incremental steps in time.

Treating the heavily-damped case, we observe that in Eq. (8), the integral in the exponent may be approximated, for  $t = t_1 \ll 1$ , to yield:

$$I_i(x, t_1) = e^{-2\sigma_i t_1} g_i(x - v_i t_1) e^{\rho_i [I_j(x, t_1) + I_j(x - v_i t_1, 0)]} \quad (9)$$

which is a nonlinear algebraic equation to solve for  $I_i(x, t)$  in terms of the initial conditions  $I_j(x - v_i t, 0)$  (at  $t = 0$  but with a shifted coordinate  $(x - v_i t_1)$ ). In principle, this nonlinear equation can be solved for any given nonlinear coupling coefficients,  $\rho_i$ ; for  $\rho_i \ll 1$ , Eq. (9) can be further approximated as:

$$I_i(x, t_1) - \rho_i e^{-2\sigma_i t_1} g_i(x - v_i t_1) I_j(x, t_1) = e^{-2\sigma_i t_1} g_i(x - v_i t_1) e^{\rho_i I_j(x - v_i t_1, 0)} \quad (10)$$

for  $i, j = 1, 2$ ;  $i \neq j$

which forms a system of coupled linear equations to be solved for  $I_i$  and  $I_j$  at  $t = t_1$ , in

terms of  $I_i$  and  $I_j$  at  $t=0$ . The solutions are:

$$I_i(x, t_1) = e^{-2\sigma_i t_1} g_i(x - v_i t_1) \frac{e^{\rho_i I_j(x - v_i t_1, 0)} - \rho_i e^{-2\sigma_j t_1} g_j(x - v_j t_1) e^{\rho_j I_i(x - v_i t_1, 0)}}{1 - \rho_i \rho_j e^{-2(\sigma_i + \sigma_j) t_1} g_i(x - v_i t_1) g_j(x - v_j t_1)} \quad (11)$$

for  $i, j = 1$  and  $2$ ;  $i \neq j$

The last equation describes a nonlinear mapping of a set of initial field distributions at  $t=0$  into another set at  $t=t_1$ . The computations of such a mapping are straightforward. After  $I_i(x, t_1)$  and  $I_j(x, t_1)$  are computed, they can be used as a new set of initial conditions for computations of the field distributions at  $t=t_2$ . This process can be repeated as the algorithm for numerical computations of the initial-value problem for given initial profiles of the waves.

Using this algorithm, we have made calculations for several examples of the three-wave interaction (heavily-damped case). One set corresponds to the Gaussian-pulse parameters used by Chu and Karney,<sup>2</sup> with results in close agreement with theirs (done by differential techniques). The results were obtained using the MRI PDP-11/40 Computational Facility and are shown in Figs. 1 to 3 using its "three-dimensional" display capability. The two interacting Gaussian pulses are labelled as:  $I_1$  and  $I_2$  with the amplitudes  $P_1$  and  $P_2$ , the half width  $W_1$  and  $W_2$  and initially at the positions  $x_1$  and  $x_2$ , respectively. The numerical results may be interpreted as follows:

ampl= 1.00000e-03 W1= 1.00000e-01 W2= 5.00000e-02  
X1= 2.00000e-01 X2= 2.00000e-01

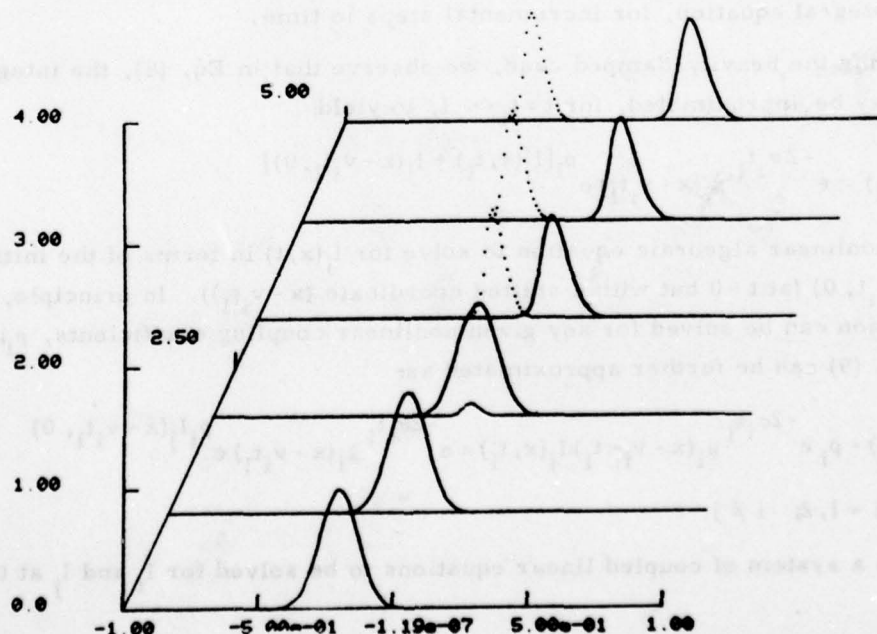


Fig. 1.

Fig. 1 (parameters --  $P_1 = 1.0$ ,  $P_2 = 10^{-3}$ ,  $W_1 = 10^{-1}$ ,  $W_2 = 0.5 \times 10^{-1}$ ,  
 $x_1 = 2 \times 10^{-1}$ ,  $x_2 = 2 \times 10^{-1}$ ):

Here the  $I_2$  pulse is too small to be seen in the first two time frames, but is incident from the right from an initial spatially-separated position. The  $I_2$  pulse has achieved sufficient gain by the third frame to be discernible (the dotted profile). By the time of the fourth time frame the  $I_2$  pulse has propagated clear of the  $I_1$  pulse, but during the intervening time the remaining overlap (interaction) has resulted in further gain of the  $I_2$  profile. Thereafter, in frames 5 and 6, both pulses separate in opposite directions with no further (perceptible) change of shape or amplitude. The total depletion of  $I_1$  by this interaction is small.

ampl= 1.00000e-02 w1= 1.00000e-01 w2= 5.00000e-02  
 x1= 2.00000e-01 x2= 2.00000e-01

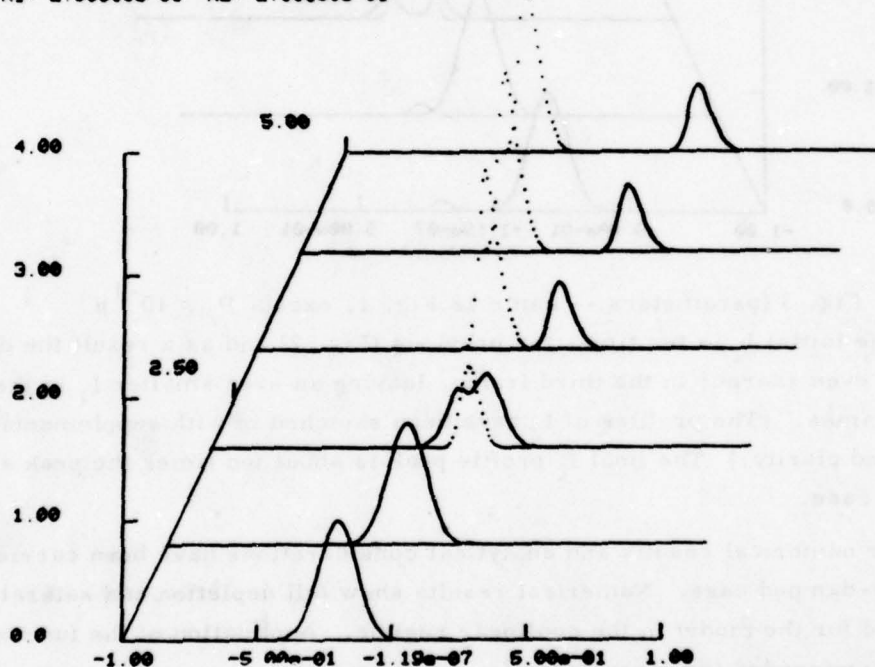


Fig. 2 (parameters -- same as Fig. 1, except  $P_2 = 10^{-2}$ ):

Here the initial  $I_2$  is ten times larger than in Figure 1. As a result, the growth of  $I_2$  is sufficiently much larger by the third time frame to cause a significant "notch" of depletion in the  $I_1$  envelope. By the time of the fourth frame, this depletion has evidently continued so as to deplete the entire trailing (left-hand) edge of  $I_1$ . The amplitude of  $I_2$ , of course, has also gained considerably more in the time lapse between frames 2 and 3, such that  $I_2$  is nearly three times the amplitude of  $I_1$ . These amplitudes and shapes again remain unchanged in the fifth and sixth frames due to no further significant interaction, as the pulses continue to separate.

ampl= 1.00000e-01 W1= 1.00000e-01 W2= 5.00000e-02  
 X1= 2.00000e-01 X2= 2.00000e-01

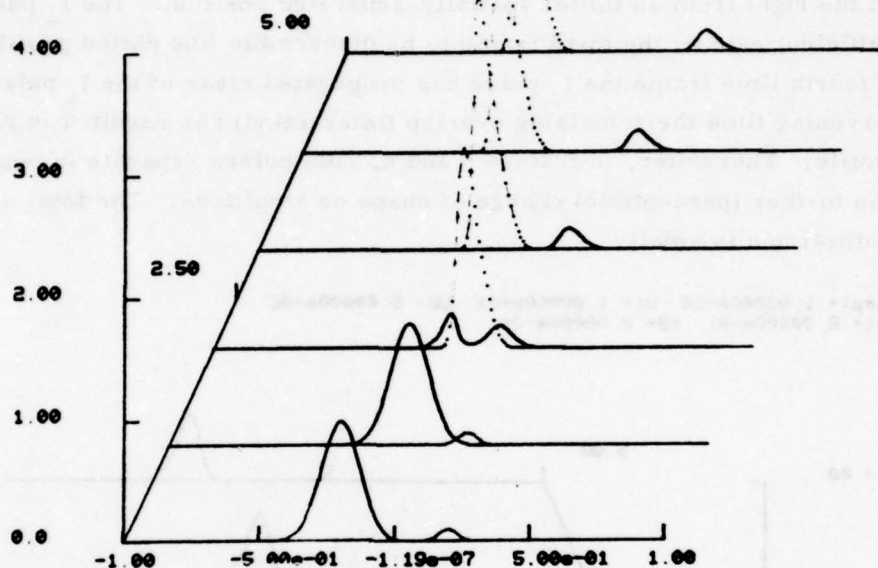


Fig. 3 (parameters -- same as Fig. 1, except  $P_2 = 10^{-1}$ ):

Here the initial  $I_2$  is ten times the previous (Fig. 2) and as a result the depletion of  $I_1$  by  $I_2$  is even sharper in the third frame, leaving an even smaller  $I_1$  in the subsequent time frames. (The profiles of  $I_2$  have been sketched in with supplementary dashed lines for added clarity.) The final  $I_2$  profile peak is about ten times the peak amplitude of  $I_1$  for this case.

Further numerical results and analytical considerations have been carried out<sup>8</sup> for the highly-damped case. Numerical results show full depletion and saturation effects expected for the model in the nonlinear regime. Application of the functional approach is recommended for other cases.

National Science Foundation  
 ENG 76-01459

E. S. Cassedy

#### REFERENCES

1. F. Y. F. Chu and A. C. Scott, Phys. Rev. **A12**, 2060 (1975).
2. F. Y. F. Chu and C. F. F. Karney, Phys. Fl. **20**, 1728 (1977).
3. R. W. Harvey and G. Schmidt, Phys. Fl. **18**, 1395 (1975).
4. A. Bers in "Plasma Physics - Les Houches," Gordon and Breach Science Publishers, 1975.

5. G.B. Whitham, "Linear and Nonlinear Waves," Wiley-Interscience, N.Y., 1974.
6. M.J. Ablowitz, et al., Stud. Appl. Math. 53, 249 (1974).
7. C.S. Lui in "Advances in Plasma Physics - Vol. 6," A. Simon and W.B. Thompson, Editors, John Wiley and Sons, N.Y. (1976).
8. E.S. Cassetty and S.T. Peng, Final Report to National Science Foundation, Grant ENG 76-01459 (March 1978).

## ANALYSIS OF STRONG ELECTROMAGNETICALLY INDUCED SPHERICALLY IMPLoding SHOCKS

Y. Fujimoto and E. A. Mishkin

A. Introduction

Spherical shock waves present solutions of the mass, momentum and energy continuity equations (1),<sup>1</sup> which are discontinuous at the shock front. They may move away from the center, either due to an increased pressure in the central domain of a star, or to a sudden release of a finite energy  $E$  at the center of the disturbance, or be center bound-imploding shocks that compress substantially the sphere of matter they bear upon. The latter can be induced by a converging "spherical piston" realized, for example, by ablated matter under the impact of a strong electromagnetic energy pulse. The increased interest in implosion is due mainly to the possibility of compressing matter to densities exceeding, by several orders of magnitude, that of solids and studying matter at very high temperatures and pressures. Also, highly compressed matter shortens, by several orders of magnitude, the mean free path of the very energetic  $\alpha$ -particles produced in a thermonuclear reaction rendering it self-sustained and, therefore, it is viewed as a necessary step in realizing controlled thermonuclear fusion.

B. Conservation Equations, Self-Similar Solution

Spherical shock waves follow the partial differential mass, momentum and energy conservation equations

$$\begin{aligned} d_t \rho + \rho \partial_r u + \frac{2}{r} \rho u &= 0 \\ d_t u + \frac{1}{\rho} \partial_r p &= 0 & \partial_r \equiv \frac{\partial}{\partial r}, \quad \partial_t \equiv \frac{\partial}{\partial t}, \\ d_t(p \rho^{-\gamma}) &= 0 & d_t = \partial_t + u \partial \end{aligned} \quad (1)$$

$p$ ,  $\rho$  and  $u$  denote the pressure, density and velocity, respectively, of the ideal gas with constant specific heats,  $\gamma = c_p/c_v$ , through which the shock wave is moving.

We denote the position of the shock front by  $R_s$  and measure the distance  $r$  of some arbitrary point behind it in units of  $R_s$ ,  $\xi = r/R_s$ . The imploding shock extends over  $1 < \xi < \infty$ . The domain  $0 < \xi < 1$ , corresponds to the wave that originated at the center. Neglecting the molecular structure of the ideal gas, the lack of any characteristic length suggests the self-similar solution,<sup>1,2</sup>

$$p(r, t) = \rho_0 \dot{R}_s^2 P(\xi); \quad \rho(r, t) = \rho_0 \dot{R}_s \rho(\xi); \quad u(r, t) = \dot{R}_s U_1(\xi), \quad (2)$$

of the conservation equations (1).  $P(\xi)$ ,  $\rho(\xi)$  and  $U_1(\xi)$  are the reduced, or non-dimensional pressure, density and velocity, respectively.

The resultant ordinary differential equations of the reduced functions are greatly simplified by the velocity transformation<sup>3</sup>

$$U_1(\xi) \rightarrow U_1(\xi) = U(\xi) + \xi, \quad (3)$$

$U$  is the speed of sound at the point  $\xi_m$  behind the shock front where the pressure is maximum.<sup>3</sup> With  $\rho_0$  constant, expressions (2) together with the transformation (3) lead to the set of ordinary differential equations in  $P$ ,  $\rho$  and  $U$ ,

$$\begin{aligned} -\rho^{-1} d_\xi \rho &= U^{-1} d_\xi U + 2\xi^{-1} + 3U^{-1}, \\ -\rho^{-1} d_\xi P &= U d_\xi U + \lambda \xi + (\lambda + 1)U, \\ -P^{-1} d_\xi P &= \gamma U^{-1} d_\xi U + 2\xi^{-1} + (3\gamma + 2\lambda)U^{-1}, \end{aligned} \quad (4)$$

where the self-similar coefficient,

$$\lambda = \frac{d \ln \dot{R}_s}{d \ln R_s} = \text{const.} \quad (5)$$

The shock front at time  $t$  is at,

$$R_s(t) = R_0 (1 + t/t_c)^\alpha, \quad \alpha(\gamma) = \frac{1}{1 - \lambda(\gamma)}. \quad (6)$$

Using the notation

$$\sigma(\xi) = \exp\left(-\int_1^\xi U^{-1} d_\xi\right), \quad \sigma(1) = 1; \quad (7)$$

$$d_\xi \sigma(\xi) = -U^{-1} \sigma(\xi),$$

we obtain the integral forms of  $P$  and  $\rho$

$$\rho(\xi) = [-\xi^2 U(\xi)]^{-1} \sigma^3(\xi); \quad (8)$$

$$P(\xi) = 2 \frac{(\gamma-1)\gamma}{(\gamma+1)^{\gamma+1}} [-\xi^2 U(\xi)]^{-\gamma} \sigma^{3\gamma+2\lambda}(\xi).$$

Eliminating  $d_\xi U$ , Eqs. (4) leads to

$$(C^2 - U^2) \frac{d_\xi P}{\xi P} = 2\gamma \left(\frac{U}{\xi}\right)^2 + (2\gamma + 2\lambda - \gamma\lambda) \frac{U}{\xi} - \gamma\lambda, \quad (9)$$

where,

$$C^2(\xi) = \gamma \frac{P(\xi)}{\rho(\xi)}; \quad c(r, t) = \dot{R}_s C(\xi)$$

$C(\xi)$  is the reduced speed of sound.

### C. Some Physical Considerations Concerning Spherical Shocks

The exploding and imploding spherical shock waves are differently generated and the underlying physical principles must reflect it. In the first case, a sudden release of a finite energy  $E$ , at the center, causes an outwardly emanating spherical shock wave. Not too far from the point of explosion the shock is still strong and the pressure which the shock wave encounters can be neglected. The energy of the mass of gas set in motion behind the shock front then is conserved or time dependent.<sup>4</sup> The total energy of the gas behind the strong shock,

$$\begin{aligned} E &= 4\pi \int_0^{R_s(t)} \left( \frac{p}{\gamma-1} + \frac{1}{2} \rho u^2 \right) r^2 dr, \\ &= 4\pi \rho_0 R_0^5 \frac{\alpha^2}{t_c^2} \left( 1 - \frac{t}{t_c} \right)^{5\alpha-2} \int_0^1 \left[ \frac{1}{\gamma-1} P(\xi) + \frac{1}{2} \lambda(\xi) (U_1 + \xi) \right] d\xi, \end{aligned} \quad (11)$$

is time independent when,

$$\alpha = 0.4 \quad (12)$$

The slopes of the reduced pressure, right behind the shock front, follow directly Eq. (9),

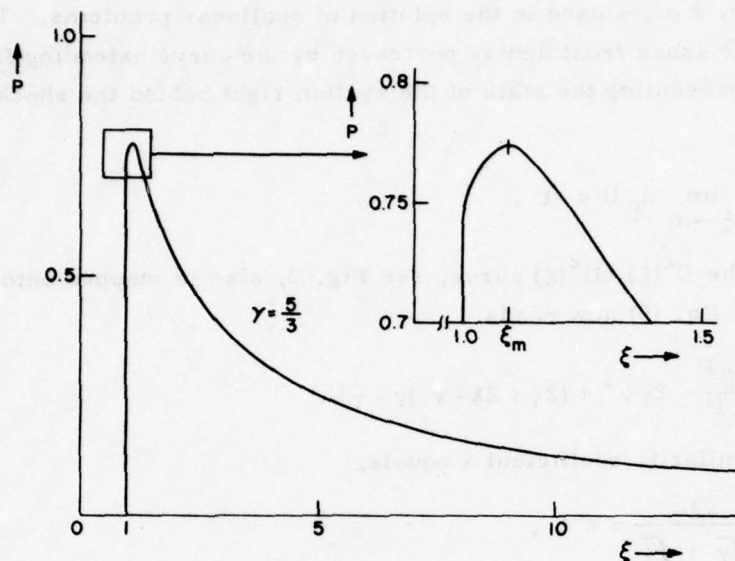
$$d_\xi P(1) = -4 \frac{2\gamma-1}{\gamma^2-1} (\lambda - \lambda_0); \quad \lambda_0 = - \frac{2\gamma(\gamma-1)}{(\gamma+1)(2\gamma-1)}, \quad (13)$$

$$d_\xi P > 0, \text{ when } \lambda < \lambda_0, \quad (14)$$

It is shown in Ref. 3 that  $\lambda$  must be smaller than, or equal to  $\lambda_0$ . In the wake of the shock the pressure  $P$  dies out. The pressure  $P$  then must reach a maximum,  $P_m$ , somewhere behind the front of the imploding shock wave. This pressure may increase slightly, reaching a maximum before it dies out at the tail of the shock. Detailed considerations show that this pressure maximum is very small and occurs very close to the shock front, see Figure 1. They also show that for  $\gamma > 2 + \sqrt{3}$ , the pressure maximum  $P_m$  occurs right at the shock front.

### D. The Postulate of Analyticity by Gelfand<sup>5</sup> and Butler<sup>6</sup>

The shock front represents a discontinuity. Behind it the pressure, the density, and the velocity are differentiable, well-behaved functions. This request for analyticity coincides, or is congruent, with the requirements of maximum pressure behind the

Fig. 1. The reduced pressure  $P(\xi)$ .

shock front. A spherically meaningful solution,  $\lambda = \lambda(\gamma)$ ,  $\alpha = 1/1 - \lambda$ , must be single valued and consider the maximum of the reduced pressure curve  $P(\xi)$ . In 1952, I.M. Gelfand showed that the function  $\lambda(\gamma)$  spans a whole interval of values and "proposed to choose the exponent on the principle of analyticity."<sup>5</sup> D.S. Butler<sup>6</sup> found  $\lambda(\gamma)$  with the aid of a high speed digital computer, he chose as he writes "... the one non-singular solution of a system of nonlinear differential equations<sup>6</sup> ....

Equations (4) lead directly to the following equation in  $\sigma$ ,

$$2 \frac{(\gamma-1)^\gamma}{(\gamma+1)^{\gamma+1}} [-\xi^2 U]^{-\gamma+1} \quad 3\gamma + 2\lambda - 3 = \frac{U d_\xi U + \lambda \xi^{-1} U + (\lambda+1)U}{\gamma U^{-1} d_\xi U + 2\gamma \xi^{-1} U + (3\gamma+2\lambda) U^{-1}} \quad (15)$$

The l.h.s. of the last equation is finite and at the singularity, when the denominator of the fraction on its r.h.s. vanishes, its numerator must vanish too. Considering again Eqs. (4) we obtain,

$$(\rho^{-1} - K P^{-1}) d_\xi P = 0 \quad (16)$$

At the singularity then the reduced pressure curve  $P(\xi)$  reaches maximum,

$$d_\xi P = 0 \quad (17)$$

#### E. The Phase-Plane $U/\xi = y$ , $dU/d\xi = x$ of the Self-Similar Solution

We introduce the  $U/\xi$ ,  $d_\xi U$  phase plant of the self-similar solution in analogy with

the phase plane  $r, \dot{r}$  often used in the solution of nonlinear problems. The motion of the system behind the shock front now is portrayed by the curve extending from the point  $U(1), d_{\xi} U(1)$ , representing the state of the system right behind the shock front, to the shock tail at,

$$\lim_{\xi \rightarrow \infty} \frac{U}{\xi} = \lim_{\xi \rightarrow \infty} d_{\xi} U = -1, \quad (18)$$

see Figure 2. The  $C^2(\xi) - U^2(\xi)$  curve, see Fig. 3, also is mapped onto the same curve. The conservation Eq. (9) now reads,

$$(C^2 - U^2) \frac{d_{\xi} P}{\xi P} = 2\gamma y^2 + (2\gamma + 2\lambda - \gamma)y - \gamma\lambda. \quad (19)$$

When the self-similarity coefficient  $\lambda$  equals,

$$\lambda = \lambda_m = \frac{-2\gamma}{(\sqrt{\gamma} + \sqrt{2})} = < 0, \quad (20)$$

The discriminant  $\Delta = 0$  and Eq. (19) reads

$$(C^2 - U^2) \frac{d_{\xi} P}{\xi P} = 2\gamma(y - y_m)^2. \quad (21)$$

Around  $y_m$ , the r.h.s. of Eq. (21) is an even function in  $y$ . In the interval,  $1 < \xi < \xi_m$ , both  $d_{\xi} P$  and  $C^2 - U^2$ , are positive. Beyond  $\xi_m$  they are both negative, see Figures 1 and 3.  $d_{\xi} P$  and  $C^2 - U^2$  on the l.h.s. of this equation vanish simultaneously at  $y_m$ , or  $\xi_m$ ,

$$d_{\xi} P(\xi_m) = 0, \quad C^2(\xi_m) - U^2(\xi_m) = 0. \quad (22)$$

At  $y = y_m$ , there is a double zero on both sides of Equation (19).

This proves the uniqueness of the  $\lambda = \lambda_m$  solution for  $1 < \gamma < 2 + \sqrt{2}$ , for a discriminant  $\Delta$  can be neither negative (there is no solution) nor positive. In the latter case there are two zeros on the r.h.s. of Eq. (21) and three on its l.h.s.

Table I shows the analytically obtained value of the self-similar coefficient  $\alpha$  and those hitherto obtained numerically.

E. A. Mishkin

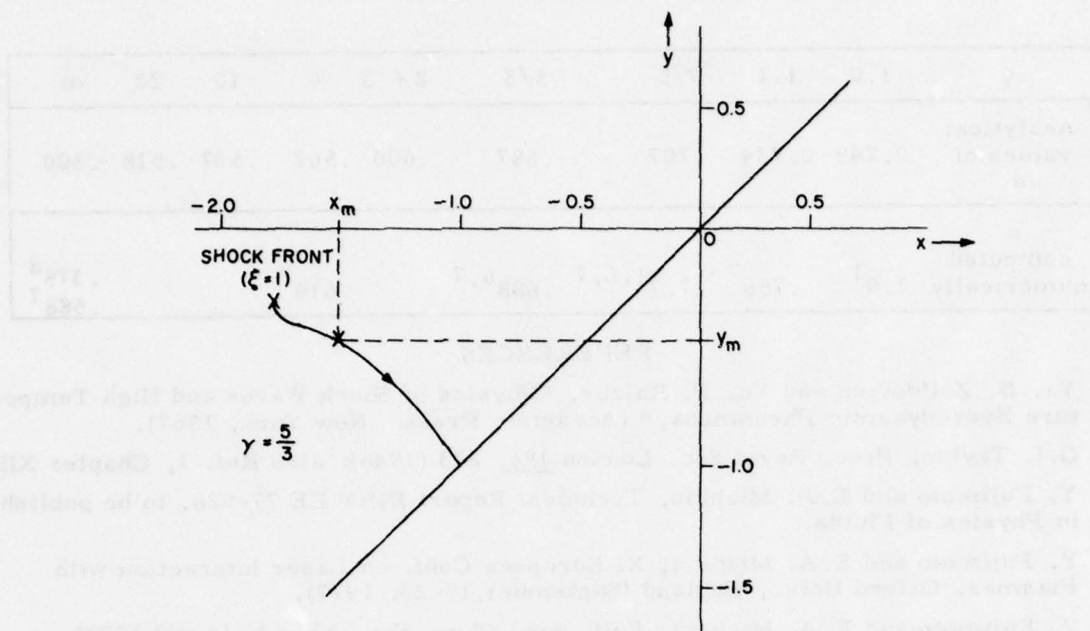


Fig. 2.

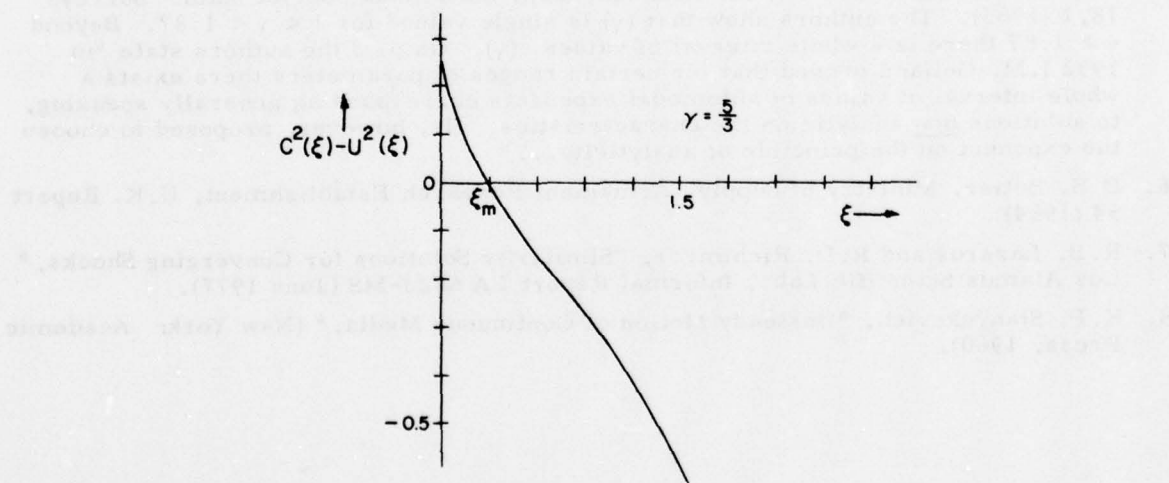


Fig. 3.

TABLE I. The self-similar exponent  $\alpha$ .

$\gamma$	1.0	1.1	7/5	5/3	2 + 3	6	10	20	$\infty$
Analytical values of $\alpha$	0.749	0.734	.707	.687	.600	.562	.537	.518	.500
$\alpha$ computed numerically	1.0 <sup>1</sup>	.786 <sup>7</sup>	.7.7 <sup>1,6,7</sup>	.688 <sup>6,7</sup>		.610 <sup>7</sup>			.375 <sup>8</sup> .588 <sup>7</sup>

## REFERENCES

1. Ya. B. Zel'dovich and Yu. P. Raizer, "Physics of Shock Waves and High Temperature Hydrodynamic Phenomena," (Academic Press: New York, 1967).
2. G.I. Taylor, Proc. Royal Soc. London 186, 273 (1946); also Ref. 1, Chapter XII.
3. Y. Fujimoto and E.A. Mishkin, Technical Report PINY EE 77-026, to be published in Physics of Fluids.  
Y. Fujimoto and E.A. Mishkin, XI European Conf. on Laser Interaction with Plasmas, Oxford Univ., England (September 19-23, 1977).  
Y. Fujimoto and E.A. Mishkin, Bull. Am. Phys. Soc. 23, 525 (April 1978).  
E.A. Mishkin and Y. Fujimoto, Technical Report PINY EE 78-043, to be published in J. Fluid Mechanics.
4. L.I. Sedov, "Similarity and Dimensional Methods in Mechanics," (New York: Academic Press, 1959).
5. K.V. Brushlinskii and Ya. M. Kazhdan, Usp. Mat. Nauk. Soviet Math. Surveys 18,1 (1963). The authors show that  $(\gamma)$  is single valued for  $1 < \gamma < 1.87$ . Beyond  $\gamma > 1.87$  there is a whole interval of values  $(\gamma)$ . On p. 3 the authors state "in 1952 I.M. Gelfand proved that for certain ranges of parameters there exists a whole interval of values of automodel exponents corresponding generally speaking, to solutions non analytic on the characteristics. He, however, proposed to choose the exponent on the principle of analyticity..."
6. D.S. Butler, Ministry of Supply, Armament Research Establishment, U.K. Report 54 (1954).
7. R.B. Lazarus and R.D. Richtmyer, "Similarity Solutions for Converging Shocks," Los Alamos Scientific Lab., Informal Report LA 6823-MS (June 1977).
8. K.P. Stanyukovich, "Unsteady Motion of Continuous Media," (New York: Academic Press, 1960).

## DEVELOPMENT OF THE DESIGN FOR IRON-CORED SYNCHRONOUSLY OPERATING LINEAR MOTORS

E. Levi

### A. Introduction

The first phase in the U.S. development of a linear propulsion system for high speed ground transportation is reaching completion. This seems to be an appropriate time to review the role played by the Polytechnic in the national program.

The Polytechnic project was initiated in 1973, after the Department of Transportation had already made two important choices with regard to linear propulsion: (1) a system in which the energy would be supplied to the vehicle, as opposed to energization by the wayside, (2) a system using iron-cored instead of superconducting magnetic.

A wheel-on-rail linear induction motor research vehicle (LIMRV) had already been tested and a similarly propelled track levitated vehicle (TLV) was under construction.<sup>1</sup> The shortcomings of the induction motor, with regard to power factor and efficiency, and the interest in attractive magnetic levitation prompted the search for alternative propulsion schemes. Favorable consideration was given to the synchronously operating motors proposed by the Polytechnic.

Consideration was restricted to magnetic field topologies which were single-sided and allowed the rail track to be completely passive. To be determined was the maximum economical clearance between the energized portion of the motor, carried by the vehicle, and the rail track. The specifications called for: thrust =  $10^5$  N, speed = 134 m/s, power = 13.4 MW, power factor = 1, efficiency  $\geq .9$ . The calculated performance parameters of the TLV LIM served as benchmarks for comparison.

### B. Project Planning

Iron-cored, synchronously operating machines with single side excitation have been used in the past, mainly as high frequency generators at the kW power level. The present task then involved transition from rotating to linear motion, and scaling-up by three orders of magnitude in power. This amounted to the development of a new machine. In this case a step by step approach led to minimal risks and costs.

The first step involved a comprehensive study of a matrix of motor types and track clearances. Next, the two most promising configurations were to be selected for detailed analysis. This would reveal the critical issues to be addressed in the next phase. Final designs for full scale prototypes and scaled-down models were to follow. As a last step, before construction of a full scale prototype, the accuracy of the performance predictions was to be tested on the scaled-down model. In the case of linear motors, this means tests on a large wheel simulating the track.

### C. Selection of the Most Promising Configurations

The simplest synchronously operating machine has a d-c excitation on one side (field), and a traveling wave a-c excitation on the other (armature). Although such a machine does not satisfy the requirement of a passive rail track, it has the simplest topology and leads to the highest power density (power output per unit weight). Since, in a transportation system, this is the most important parameter, a design was performed for the conventional synchronous machine and was used as reference base.

Attention was focused on three machine types in which both field and armature windings are carried by the vehicle side. These were: (1) the homopolar inductor first proposed by Kemper<sup>2</sup> and later analyzed by Rummich<sup>3</sup> (Fig. 1), the claw-pole or Nadyne first proposed by Pierro<sup>4</sup> (Fig. 2), and the heteropolar inductor machine<sup>5</sup> (Figure 3).

In the homopolar inductor machine, the field winding is interlinked with a transverse d-c flux; the armature winding is divided into two sections, each interlinked with both a longitudinal a-c flux and the transverse d-c flux. As a result, in the armatures the magnetic flux density  $B$  oscillates around an average value, instead of alternating between a positive and a negative peak, as in conventional machines.

The same occurs in the heteropolar inductor machine, even though both field and armature fluxes are longitudinal.

In the claw-pole machine, instead, the magnetic flux path is twisted from transverse to longitudinal, while it is forced to traverse two additional air gaps.

Two clearance lengths (1.5 cm and 3 cm) were considered. The heteropolar inductor machine carries an alternating flux through the rail and hence, requires a finely laminated rail track. Therefore, it was not deemed suited for ground rail transportation. This left the homopolar inductor and the claw-pole types as sole contenders. Motivation for pursuing their development was provided by the fact that both, according to the analysis, outperformed the TLV LIM.<sup>6</sup>

With regard to the homopolar inductor type it must be noted that, due to its asymmetry, the flux swing and, hence, the output is severely limited by iron saturation effects. This limitation is compounded by the fact that saturation also hinders the variable reluctance effects on which the machine relies for its operation.

On the other hand, in the claw-pole machine, the additional mmf required to overcome the double air gap enhances the role of leakage fluxes. As it turns out, the weight penalty imposed by these fluxes limits the usefulness of the machine to clearances well below 1 cm. For this reason, further development was concentrated on the inductor motor.

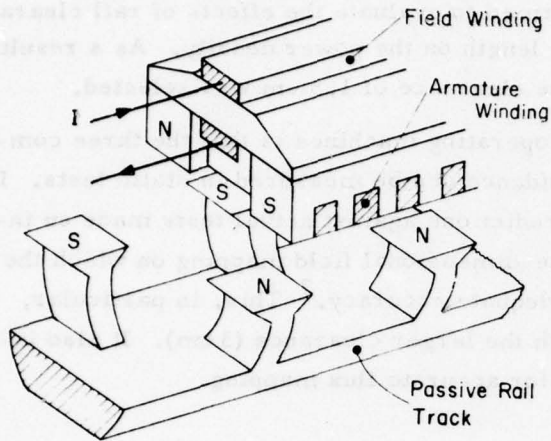


Fig. 1. Sketch of homopolar inductor motor.

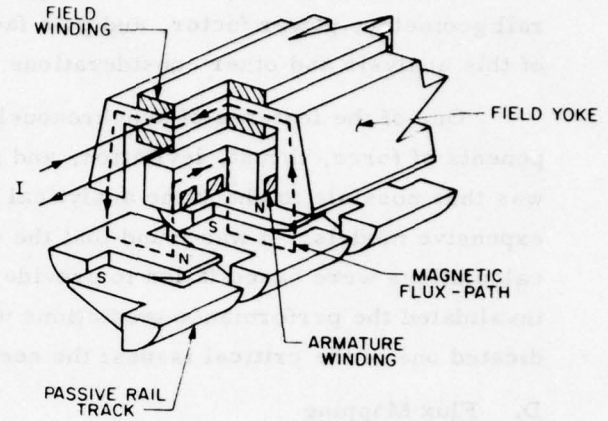


Fig. 2. Sketch of claw-pole motor.

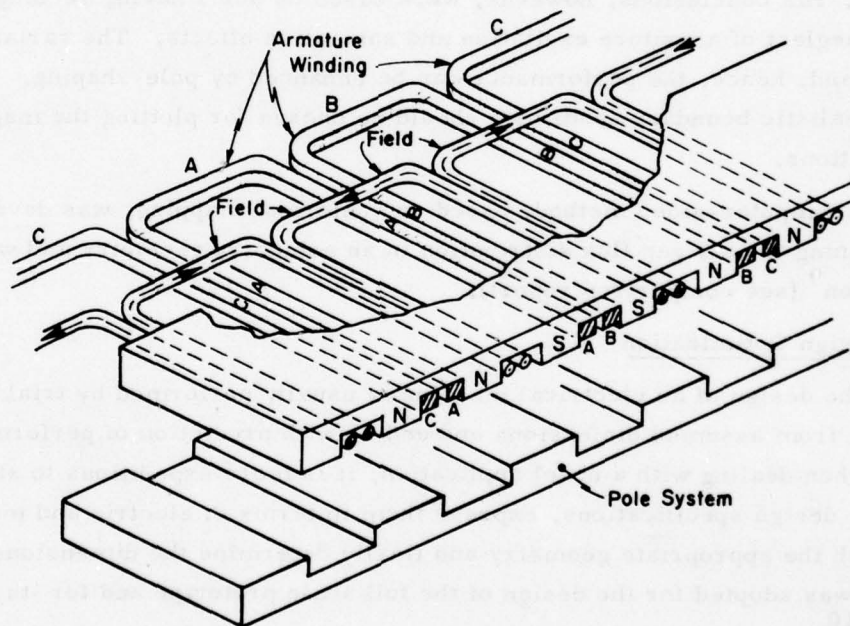


Fig. 3. Heteropolar inductor motor.

A sensitivity analysis was then performed to evaluate the effects of rail clearance, rail geometry, power factor, and pole face length on the power density. As a result of this analysis and other considerations the clearance of 1.5 cm was selected.

One of the features of synchronously operating machines is that the three components of force, thrust, levitation, and guidance can be measured in static tests. It was thus possible to check the analytical predictions against actual tests made on inexpensive models. It was found that the one-dimensional field mapping on which the calculations were based failed to provide adequate accuracy.<sup>7</sup> This, in particular, invalidated the performance predictions with the larger clearance (3 cm). It also indicated one of the critical issues: the need for accurate flux mapping.

#### D. Flux Mapping

Inductor machines were studied at the beginning of the century by Wieseman.<sup>8</sup> Using graphical flux plotting techniques, he determined the optimum pole arc to be  $\alpha = 60^\circ$ . His conclusions, however, were based on poles having rectangular profiles and on neglect of armature excitation and saturation effects. The variable reluctance effects and, hence, the performance can be enhanced by pole-shaping. Moreover, more realistic boundary conditions should be chosen for plotting the magnetic flux distributions.

A computer-aided method, based on conformal mapping, was developed for determining the air gap flux distribution in an arbitrary geometry and with arbitrary excitation<sup>9</sup> (see comparison report).

#### E. Design Optimization

The design of an electrical machine is usually performed by trial and error, starting from assumed dimensions and ending with prediction of performance. However, when dealing with a novel application, it is more expeditious to start directly with the design specifications, express them in terms of electric and magnetic stresses, establish the appropriate geometry and finally determine the dimensions. This approach was adopted for the design of the full scale prototype and for its scaled-down model.<sup>10</sup>

The main design premises were:

- (a) iron saturation, rather than thermal stress, is the basic limitation of the machine.
- (b) supply by a current-source inverter under closed loop operation allows selection of the angle  $\psi$  between the direct axis and the armature current maximum.

It was found that, for unity power factor, there exists an optimum value  $\psi \approx \alpha \approx 50^\circ$ . This choice of  $\psi$  is characterized by a significant reduction of the additive effects of the armature reaction on the air gap flux density under the pole faces. In turn this reduces saturation of the armature teeth, the most stressed part of the iron. In addition, at the same time, this choice of  $\psi$  maximizes the thrust per unit gap area and, hence, for a specified power output, tends to lead to a machine of minimum size.

#### F. End Effects

One of the advantages of linear synchronous machines is that they do not rely for their operation on induced currents. Therefore, the end effects which seriously curtail the output of linear induction motors can be alleviated. Nevertheless, it was necessary to evaluate the undesirable finite length effects of the homopolar inductor motor.<sup>11</sup> It was found that the permeance fluctuations cause negligible losses. More important is the possible development of unbalanced forces. Undesirable yaw, pitch, and roll motions may be minimized by making the overall length of the motor equal to an integer number of pole pairs.

#### G. Laminations of the Rail Track

The question of whether or not the rail track should be laminated and, if so, what should be the lamination thickness has important economic implications. The problem does not easily lend itself to a rigorous analysis. It was, therefore, investigated following four different approaches.<sup>11</sup> From these studies, it appears that eddy currents, if allowed to flow unhindered in the rail, effectively prevent the establishment of a flux over the leading portion of the motor. Hence, the use of a solid rail is not feasible and laminations are necessary. An approximate analysis, which accounts for saturation effects, leads to the conclusion that the laminations need not be wider than 6 mm. This thickness is believed to be structurally adequate, economically feasible and leads to negligible losses. A rigorous analysis, based on the periodic structure approach is in progress.

#### H. Conclusions

An investigation of various types of iron-cored, synchronously operating linear motors has shown that for clearances in excess of 1 cm., the homopolar inductor type is the most promising linear motor. This conclusion, as well as the accuracy of the analysis, has been confirmed by a comparative evaluation of the linear inductor and induction motors carried out by General Electric on 112-kW models.<sup>12</sup> Further confirmation is provided by models constructed at the University of Kentucky,<sup>13</sup> University of Toronto,<sup>14</sup> and the Technical University of Munich.<sup>15</sup>

REFERENCES

1. M. Guarino, Jr., "Integrated Linear Electric Propulsion Systems for High Speed Transportation," Symposium International sur les Moteurs Electriques Linéaires, Lyon, France (May 1974).
2. H. Kemper, "Elektrisch Angetriebene Eisenbahnfahrzeuge mit Elektromagnetischer Schwebefuehrung," ETZ pp. 11-14 (1953).
3. E. Rummich, "Machines Linéaires Synchrones, théorie et réalisation pratiques," Bull, ASE, Vol. 63, No. 23, pp. 1338-44 (November 1972).
4. J. J. Pierro, U.S. Patent No. 3,456,136 July 15, 1969.
5. E. Levi, "Linear Synchronous Motors for High-Speed Ground Transportation, IEEE Trans. Vol. MAG-9, No. 3, pp. 242-8 (September 1973).
6. E. Levi, "High Speed, Iron-cored, Synchronously Operating Linear Motors," Proc. of the IEEE Conf. on Linear Electric Machines, pp. 155-160, London (October 1974).
7. E. Levi, "A Preliminary Evaluation of Electrical Propulsion by Means of Iron-cored Synchronously Operating Motors," NTIS No. PB-258437 (January 1975).
8. R. W. Wieseman, "Graphical Determination of Magnetic Fields," Trans. of the AIEE, Vol. 4, No. 2, pp. 141-154 (February 1927).
9. E. Levi, J.P. Lee, F. Lalezari and M. Gemelos, "Computer-aided Conformal Mapping of Magnetic Fluxes in Saturated Inductor Motors," IEEE Trans., Vol. Mag-14, No. 5, pp. 927-929 (September 1978).
10. E. Levi, L. Birenbaum and Z. Zabar, "Concerning the Design of Inductor Synchronous Motors Fed by Current Source Inverters," IEEE Trans. Vol. MAG-13, No. 5, pp. 1421-3 (September 1977).
11. E. Levi, "Preliminary Design Studies on Iron-cored Synchronously Operating Linear Motors," Polytechnic EE/EP Report No. 76-005 (February 1976).
12. T. R. Haller, "A Comparison of Linear Induction Motors and Linear Synchronous Motors for High Speed Ground Transportation," Proc. Intermag. Conference, Florence, Italy (May 9-12, 1978).
13. I. Boldea and S. A. Nasar, "Fields Winding Drag and Normal Forces of Linear Synchronous Homopolar Motors," Electric Machines and Electromechanics, pp. 253-268 (1978).
14. J. Mukolera and G. R. Slemon, "A Homopolar Linear Synchronous Motor," International Conference on Electric Machines, Brussels (September 1978).
15. H. W. Lorenzen and W. Wild, "The Synchronous Linear-Motor (in German) Report: The Technical University of Munich, West Germany (1976).

## COMPUTER-AIDED CONFORMAL MAPPING OF MAGNETIC FLUXES IN SATURATED INDUCTOR MOTORS

E. Levi, J. P. Lee, F. Lalezari and M. Gemelos

A. Introduction

A simple approach, developed for the design of linear motors, leads to a quick evaluation of the machine reactances and of their dependence on iron saturation. The method consists of conformal mapping by means of the Schwarz-Christoffel transformation<sup>1</sup> coupled with numerical integration. This allows the handling not only of a large number of pole shapes, but also of variable potential distributions at the boundaries.

B. Salient-Pole Shaping

An inductor motor for linear electric propulsion is sketched in Figure 1. Both the field and armature windings are located in the same structure, while the rail track

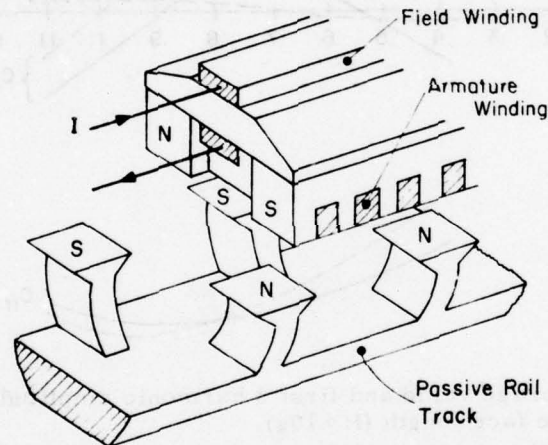
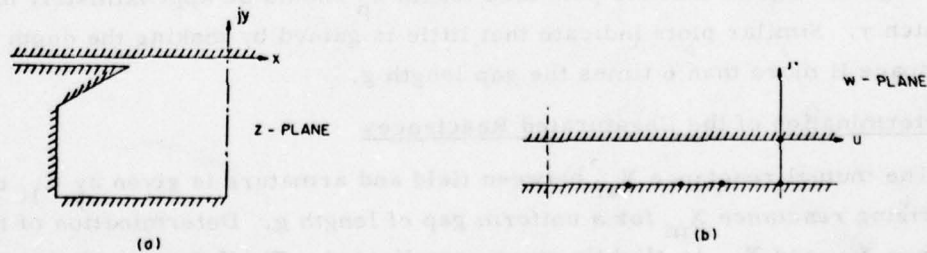


Fig. 1. Sketch of homopolar inductor motor.

is passive. Since the machine relies for its operation on variable reluctance effects, the primary design choice is a pole profile which accentuates the variation in reluctance. This effect is most simply studied using field excitation only. A two-step Schwarz-Christoffel transformation reduces the idealized geometry of Fig. 2(a) to the parallel plate configuration of Figure 2(b). Plots, such as shown in Fig. 3, giving the harmonic

Fig. 2. Salient pole profile.  
(a) Original (b) Transformed.

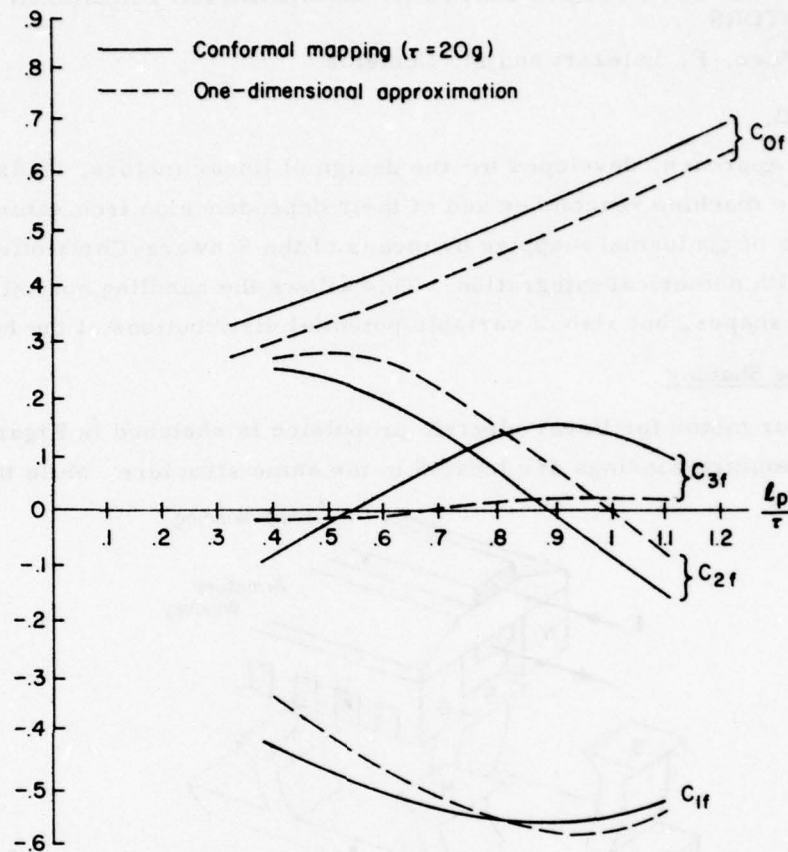


Fig. 3. Average value and first 3 harmonic amplitudes vs. pole face length ( $H = 10g$ ).

coefficients  $C_f$  (amplitude  $B_n$  of the flux density harmonic content normalized to the peak field) guide the choice of pole shapes and dimensions. For instance,  $B_0$  determines the cross section of the field yoke and of the passive rail track and, hence, affects the weight and cost of the propulsion system. In contrast, for a given armature area, surface current density, and power factor,  $B_1$  is proportional to the machine output. Hence  $C_{1f}/C_{0f}$  is an index of the output/weight ratio and a figure of merit for the system design. Figure 3 shows that the pole face length  $\ell_p$  should be approximately half of the pole pitch  $\tau$ . Similar plots indicate that little is gained by making the depth of the interpolar space  $H$  more than 6 times the gap length  $g$ .

#### C. Determination of the Unsaturated Reactances

The mutual reactance  $X_{af}$  between field and armature is given by  $C_{1f}$  times the magnetizing reactance  $X_m$  for a uniform gap of length  $g$ . Determination of the armature reactance  $X_{ad}$  and  $X_{aq}$  is slightly more complicated. The Schwarz-Christoffel trans-

formation distorts the armature boundary. Consequently the magnetostatic potential distribution along the armature is deformed and must be calculated, before it is applied to the parallel plate configuration. However the field distribution can be determined analytically and is easily transformed back into the original geometry. The resultant reactance coefficients are shown in Figure 4. For the sake of comparison, Figs. 3 and 4 also show the values of the reactance coefficients obtained by the conventional 1-dimensional (straight-line) approximation.

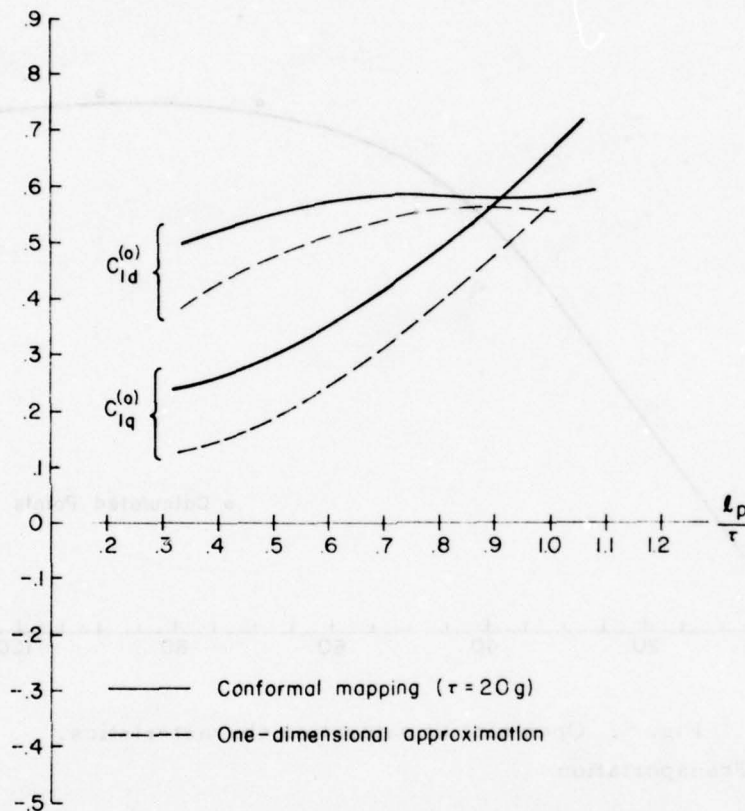


Fig. 4. Direct and quadrature coefficients vs. pole face length ( $H = 10g$ ).

#### D. Correction for Iron Saturation

The same technique is applied to account for saturation effects. The magnetic flux in the air region resulting from the combined field and armature excitations is subdivided into tubes of equal flux for which the computer calculates the coordinates along the iron. The corresponding m.m.f. drops are evaluated with the help of the iron magnetization curve. The deviations in magnetostatic potential resulting from these drops are then treated as additional excitations. Superposition applies, since the air region is linear. A first order correction is thus introduced in the flux distribution. The process can be iterated, if more accuracy is needed.

### E. Comparison with Experiment

A homopolar inductor linear motor with nominal output of 112 kW was built by General Electric and tested on a 1.35 m diameter wheel.<sup>2</sup> The 60 Hz open-circuit saturation curve data obtained during one of these tests is shown as the solid curve in Figure 5. Predictions based on the theory described earlier are shown for comparison as 3 calculated points in the figure. Excellent agreement is apparent.

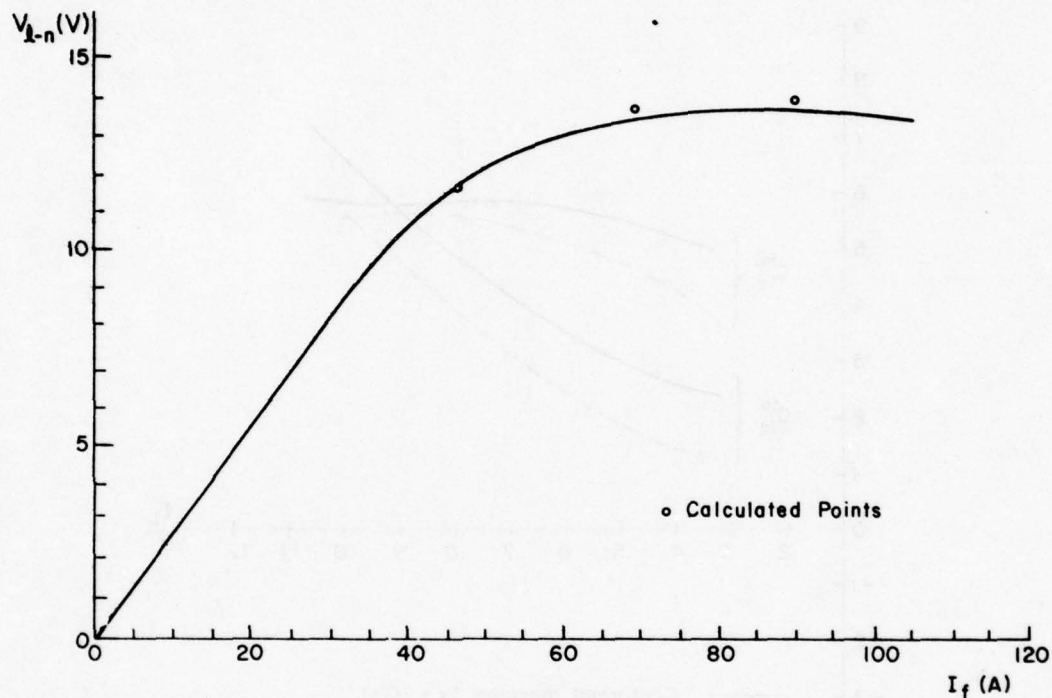


Fig. 5. Open circuit saturation characteristics.

Department of Transportation  
DOT-FR-30030  
DOT-FR-64227

E. Levi

### REFERENCES

1. L.V. Bewley, "Two-Dimensional Fields in Electrical Engineering," MacMillan (1948).
2. T.R. Haller, "A Comparison of Linear Induction Motors and Linear Synchronous Motors for High-Speed Ground Transportation," Proc. Intermag Conf., Florence, Italy (May 9-12, 1978).

# MODAL REPRESENTATION OF E-M FIELDS IN LAMINATED MOVING CONDUCTIVE FERROMAGNETIC MEDIA

B.R. Cheo, E. Levi and K.C. Chang

## A. Introduction

Early in the development of electrical machines, some 150 years ago, it was realized that the portions of magnetic flux path carrying an ac flux need to be laminated in order to reduce the effects of eddy currents. Yet the accurate evaluation of the effects is still an unresolved problem. The difficulty stems from two sources: geometrical complexity and iron saturation. Most of the engineering purposes have been served by approximations which are valid in the case of thin laminations.<sup>1-7</sup> Technological progress in materials and manufacturing have made the use of such thin laminations economically feasible.

However, in applications of the recently developed interests in magnetic propulsion of vehicles, the track rail itself has become part of the magnetic circuit. Economic and structural considerations have made it essential to use the coarsest possible laminations.

This report addresses itself to the first step of the classical problem of thick laminations: the modal structure of EM fields in a laminated moving conductive ferromagnetic medium and the boundary value problem associated with the excitation in an air gap.

## B. Formulation and Coordinate System

We consider, according to Fig. 1, a half space of laminated structure occupying

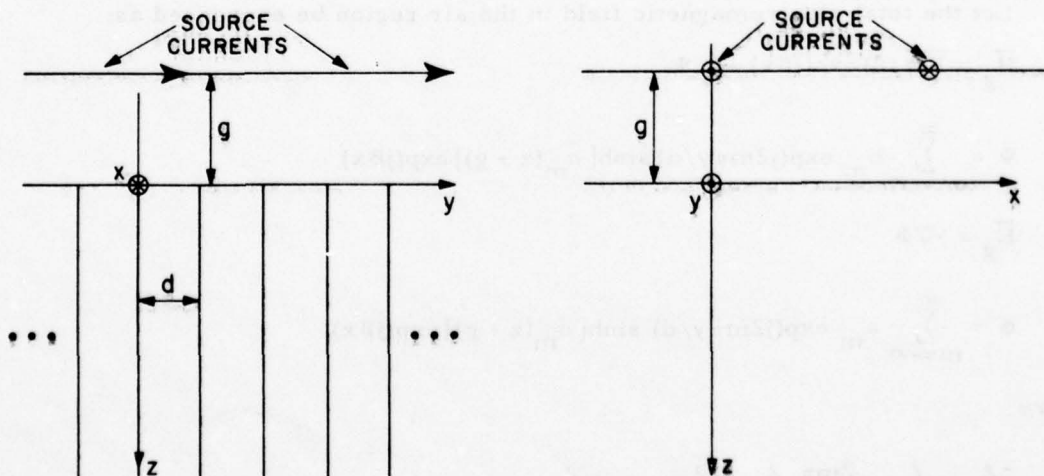


Fig. 1

the region  $z > 0$ . The medium moves with a constant velocity  $v = \vec{a}_x v$ . The static stationary sources are confined in the region  $z < 0$ . For simplicity we assume that they are uniform in the  $y$  direction, normal to the lamination sheets.

The analysis is carried out in a coordinate system fixed to the sources. Since the structure seen by the sources is unvarying, time dependence is eliminated from the problem.

### C. Fields in the Air Gap

We assume that in the absence of the lamination structure, the sources produce a known source field  $\vec{H}_s(x, z)$  that is expressible as a Fourier integral:

$$\vec{H}_s(x, z) = \int_{-\infty}^{\infty} \vec{F}(z, \beta) \exp(j\beta x) d\beta \quad (1)$$

where

$$\vec{F}(z, \beta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \vec{H}_s(x, z) \exp(-j\beta x) dx \quad (2)$$

In the presence of the moving lamination structure, an additional set of electromagnetic fields will be induced within the lamination and in the air so that the boundary conditions at the interface  $z = 0$  are satisfied. We first establish the modal form of the induced fields due to a source field  $\vec{F}(z, \beta) \exp(j\beta x)$ . Both  $\vec{E}$  and  $\vec{H}$  fields will, in general, be present and each can be expressed as the gradient of a scalar potential. They all have  $\exp(j\beta x)$  dependence and the true fields are obtained through the Fourier transform in the form of Equation (1).

Let the total electromagnetic field in the air region be expressed as:

$$\vec{H}_g = \vec{F}(z, \beta) \exp(j\beta x) - \nabla \Psi \quad (3)$$

$$\Psi = \sum_{m=-\infty}^{\infty} b_m \exp(j2m\pi y/d) \sinh[\hat{\alpha}_m(z+g)] \exp(j\beta x)$$

$$\vec{E}_g = -\nabla \Phi \quad (4)$$

$$\Phi = \sum_{m=-\infty}^{\infty} a_m \exp(j2m\pi y/d) \sinh[\hat{\alpha}_m(z+g)] \exp(j\beta x)$$

where

$$\hat{\alpha}_m^2 = \beta^2 + \left(\frac{2m\pi}{d}\right)^2, \quad \nabla^2 \Phi = 0, \quad \nabla^2 \Psi = 0$$

The hyperbolic sine  $\sinh[\alpha_m(z+g)]$  dependences in the magnetic potential  $\Psi$  and the electric potential  $\Phi$  are chosen according to the symmetry of the geometry. This choice implies the assumption that  $z = -g$  is an equi-potential surface for both, and is taken to be the reference. The coefficients  $a_m$  and  $b_m$  are to be determined.

#### D. Modes in the Laminations

In the iron, we seek a set of modes which are solutions to the Galilean transformed Maxwell's equations and satisfy the boundary conditions at the edge of laminations  $y = 0, \pm d, \pm 2d, \dots$ . The solution will be periodic in the  $y$ -direction with periodicity  $d$ .

The Galilean transformed Maxwell's equations considered are:

$$\nabla \times \vec{E} = 0 \quad (5)$$

$$\nabla \times \vec{H} = \vec{J} = \sigma(\vec{E} + \mu \vec{v} \times \vec{H}) \quad (6)$$

$$\nabla \cdot \vec{H} = 0 \quad (7)$$

where  $\vec{v} = \vec{a}_x v$ ,  $\sigma$  and  $\mu$  are the velocity conductivity and permeability of the moving medium. The fields  $\vec{E}$ ,  $\vec{H}$  here are those seen by the stationary observer, i.e., the observer at rest with respect to the source currents. The laminations are separated by infinitesimally thin perfect insulations. The boundary conditions are:

- (1) All components of  $\vec{H}$  are continuous.
- (2) Normal component of  $\vec{J} = 0$  ( $J_y = 0$ ) at the edge of the laminations ( $y = 0, \pm d, \pm 2d, \dots$ ).
- (3) A finite voltage may exist across the insulation.

All fields are assumed to have  $\exp(j\beta x)$  behavior as stated previously.

It can be shown that there are two types of modes which exist in the medium satisfying the above: The first type is the TM modes which consist of both  $\vec{E}$  and  $\vec{H}$  fields. The second type, the TE modes, consists of the  $\vec{H}$  field only.

It can be shown that a complete representation of the total fields is given by the following equations.<sup>8</sup>

$$E_x = \frac{j\beta E_o}{\gamma} \sinh\left[\gamma\left(y - \frac{d}{2}\right)\right] \exp(-\alpha z) + \sum_{n=1}^{\infty} E_n \left(\frac{-j\beta d}{n\pi}\right) \cos\left(\frac{n\pi y}{d}\right) \exp(-\alpha_n z) \quad (8)$$

$$E_y = E_o \left\{ \cosh\left[\gamma\left(y - \frac{d}{2}\right)\right] - \frac{2}{\gamma} \sinh\left(\frac{\gamma d}{2}\right) \delta(y-d) \right\} \exp(-\alpha z) \\ + \sum_{n=1}^{\infty} E_n \left[ \sin\left(\frac{n\pi y}{d}\right) - \frac{2d}{n\pi} \epsilon_n \delta(y-d) \right] \exp(-\alpha_n z) \quad (9)$$

$$E_z = \frac{-\alpha E_0}{\gamma} \sinh\left[\gamma\left(y - \frac{d}{2}\right)\right] \exp(-\alpha z) + \sum_{n=1}^{\infty} E_n \left(\frac{\alpha_n d}{n\pi}\right) \cos \frac{n\pi y}{d} \exp(-\alpha_n z) \quad (10)$$

$$H_x = \frac{\alpha \sigma}{\gamma} E_0 \cosh\left[\gamma\left(y - \frac{d}{2}\right)\right] \exp(-\alpha z) + \sum_{n=1}^{\infty} \left[ (j\beta - \sigma \mu \nu) \frac{d}{n\pi} \epsilon_n H_n - \alpha_n \sigma \left(\frac{d}{n\pi}\right)^2 E_n \right] \sin\left(\frac{n\pi y}{d}\right) \exp(-\alpha_n z) \quad (11)$$

$$H_y = \sum_{n=1}^{\infty} H_n \epsilon_{n+1} \cos \frac{n\pi y}{d} \exp(-\alpha_n z) \quad (12)$$

$$H_z = \frac{j\beta \sigma}{\gamma} E_0 \cosh\left[\gamma\left(y - \frac{d}{2}\right)\right] \exp(-\alpha z) + \sum_{n=1}^{\infty} \left[ -\frac{\alpha_n d}{n\pi} \epsilon_{n+1} H_n - j\beta \sigma \left(\frac{d}{n\pi}\right)^2 E_n \right] \sin \frac{n\pi y}{d} \exp(-\alpha_n z) \quad (13)$$

where:

$E_n$ : modal amplitudes of the TM modes

$H_n$ : modal amplitudes of the TE modes

$\epsilon_j = 1$  for  $j = \text{odd}$  and  $\epsilon_j = 0$  for  $j = \text{even}$

$\alpha = |\beta|$

$\gamma = (j\beta \mu \nu)^{1/2}$

$\alpha_n = [\beta^2 + (\frac{n\pi}{d})^2 + j\beta \sigma \mu \nu]^{1/2}$

The  $e(j\beta x)$  dependence is understood throughout. The expressions above are valid for the unit cell:  $0^+ < y < d^+$ , i.e., for the region including the first lamination and the first insulation sheet to the right of  $x$ - $z$  plane. The fields extend from  $y = -\infty$  to  $y = \infty$  periodically from cell to cell with periodicity  $d$ . The  $\delta$ -function gives rise to a finite voltage between the laminations and is needed for  $\oint \vec{E} \cdot d\vec{l} = 0$  everywhere.

The terms corresponding to  $E_0$  in Eqs. (8) to (13) are those of the dominant mode. The identity  $\alpha = |\beta|$  indicates that this mode has the least attenuation in the  $z$ -direction and thus corresponds to the deepest penetration into the iron from the air gap. It can also be shown that  $J_y = 0$  everywhere for this mode. The other terms ( $E_n$  and  $H_n$ ) correspond to the higher order modes. They all attenuate very rapidly with increasing distance from the air gap. In the limit of  $d \rightarrow 0$ , only the dominant mode survives and only  $E_y$ ,  $H_x$  and  $H_z$  remain. All other components vanish as well as the current. The medium thus behaves as if  $\sigma = 0$ .

### E. The Boundary Value Problem

After establishing the modes in the air gap and in the iron laminations, we outline the procedure for solving the modal amplitudes  $a_m$ ,  $b_m$ ,  $E_n$  and  $H_n$ . The procedure is as in the standard scattering problem of a periodic grating; i.e., by matching the total fields in the air and in the iron at the interface  $z=0$ . The following boundary conditions serve as the basis:

- (1) The tangential components of  $\vec{E}$  and  $\vec{H}$  (x and y components) must be continuous.
- (2) The normal component (z-component) of  $\vec{B}$  must be continuous.
- (3) The normal component of the current density  $\vec{J}$  must vanish.

We note that in a wave scattering problem, condition (1) will be sufficient to determine the modal amplitudes uniquely. Since we are dealing with the time invariant Galilean transformed systems, the normal component continuity also needs to be imposed. The three conditions above are not independent of each other. Using the first two, the third will be satisfied automatically. However, it is often judicious to use all three in combinations. A great deal of simplification of the algebra may result.

To date we have looked into the problem of the fields excited by a pair of line current  $\pm I_0$  located at  $z=-g$ ,  $x=0$  and  $x=L$  respectively. The current at  $x=0$  flows in the  $+y$  direction as shown in Figure 1. We consider first the case for one line current at  $x=0$ . Then

$$F(z, \beta) = (a_x - j \operatorname{sgn} \beta a_z) I_0 \exp[-|\beta|(z+g)] \quad (14)$$

where

$$\operatorname{sgn} \beta = \begin{cases} 1 & \text{for } \beta > 0 \\ -1 & \text{for } \beta < 0 \end{cases}$$

To establish conditions (1) to (3), we first recall that the expressions (8) to (13) are descriptions of the fields in the first unit cell. At  $z=0$ , these expressions can be expressed as a Fourier series in  $y$  (periodicity  $d$ ) of the form  $\sum f_m \exp(j 2n\pi y/d)$ . By matching the coefficients with those given by Eqs. (3) and (4), we obtain several sets of infinite-by-infinite algebraic equations for the unknowns  $(a_m, b_m)$  and  $(E_n, H_n)$ . After much algebra, the following results are obtained:

$$E_n = - \frac{\gamma^2 I_0 \pi^3 \exp(-2|\beta|g)}{8\sigma[\alpha_n \cosh(|\beta|g) + \mu_r |\beta| \sinh(|\beta|g)](n^2 \pi^2 + \gamma^2 d^2)} \quad (15)$$

$$\mu_r = \mu/\mu_0$$

$$E_o = \frac{\gamma d (\gamma^2 d^2 + 4\pi^2)}{\sinh(\gamma d/2)\pi^3} \sum_{n=\text{odd}} \frac{E_n}{n(n^2 - 4)} \quad (16)$$

$$b_o = \frac{1}{j\beta \sinh(\alpha_o g)} \left\{ I_o \exp(-|\beta|g) + \sum_{n=\text{odd}} \frac{2\sigma \alpha_n d^2}{(n\pi)^3} \left[ 1 - \frac{\alpha}{\alpha_n} \left( 1 - \frac{4\pi^2}{\gamma^2 d^2} \right) \left( \frac{n^2}{n^2 - 4} \right) \right] E_n \right\} \quad (17)$$

$$H_n = 0, E_n = 0 \text{ for } n \text{ even, } a_m = 0, b_m = 0 \text{ for } m > 0 \quad (18)$$

Where  $g_n$  satisfies

$$\sum_{n=\text{odd}} \frac{g_n}{(n^2 - 4m^2)} = \frac{1}{m^2}, \quad m = 1, 2, 3, \dots \quad (19)$$

we notice the following significant results:

- (1) From Eq. (18) we see that there are no TE modes in the iron.
- (2) Only odd TE modes exist.
- (3) There is no electric field in the air gap.
- (4) The magnetic field in the air gap is smooth in  $y$ .

Finally, from the computational point of view, the most significant result is given by Eqs. (15) and (19); Eq. (15) shows that the modal amplitudes  $E_n$  are equal to a known function containing all the parameters ( $\alpha, g, \sigma, v, \mu$  and  $\beta$ ) multiplied by  $g_n$ ; and Eq. (19) shows that  $g_n$  is a function of  $n$  only. Therefore, the only numerical work here is the solution of the infinite set of Eqs. (19) once. The modal amplitudes for all cases are then determined. We have solved Eq. (19) numerically, and found the values of  $g_n$  for all  $n < 203$ . It is of great interest to note that, within the numerical accuracy of the computation,  $g_n$  behaves as  $cn^{-2}$  with  $c \approx -3.24$ . This remarkable behavior is shown in Fig. 2 where  $g_n$  is plotted over four decades on a log-log graph, given along with the  $n^{-2}$  straight line for comparison.

U.S. Department of Transportation  
DOT-FR-64227

B.R. Cheo

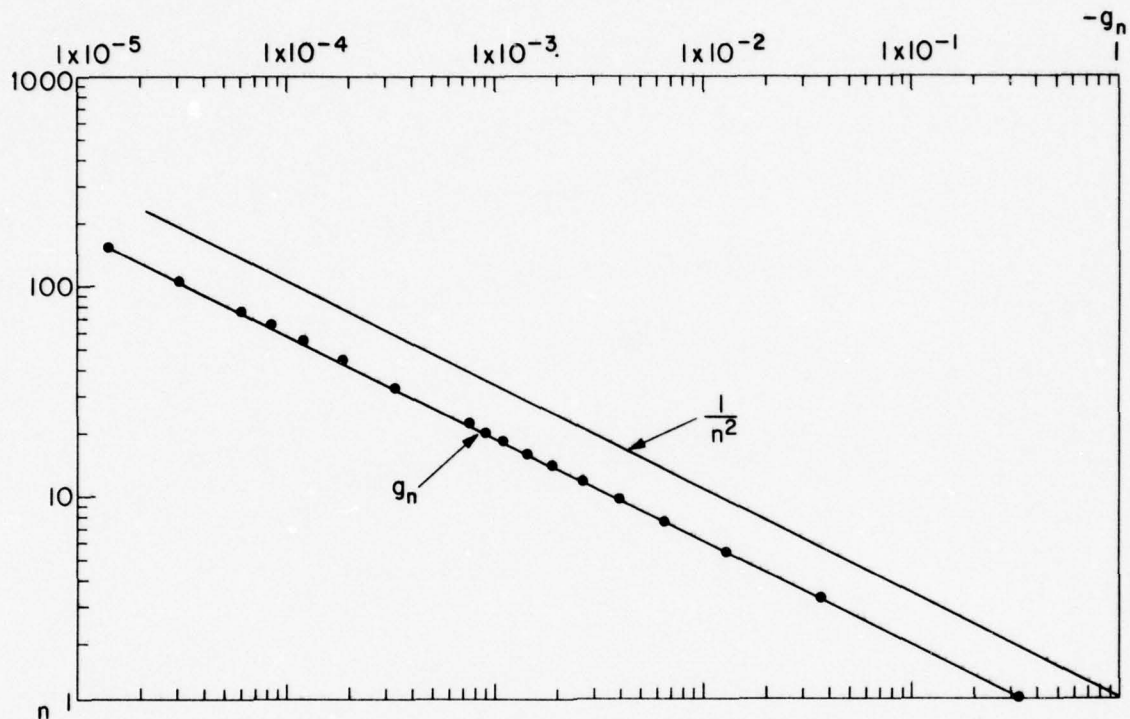


Fig. 2

## REFERENCES

1. L. Dreyfus, "Feldverteilung und Wirbelstrombildung in Dynamoankern," Arch. f.E., 4, pp. 99-139 (1915).
2. E. Rosenberg, "Wirbelströme in Massive Eisen," Elektrotechnik und Maschinenbau, Vienna, 41, pp. 317-325 (June 1923).
3. F. Ollendorff, "Hysteresis und Wirbelströme in Eisenblechen," Arch. f.E., 14, p. 431 (1925).
4. T. Spooner, "Properties and Testing of Magnetic Materials," (New York: McGraw Hill, 1927).
5. W. MacLean, "Theory of Strong Electromagnetic Waves in Massive Iron," J. Appl. Phys., Vol. 25, pp. 1267-1270 (October 1954).
6. P.D. Agarwal, "Eddy Current Losses in Solid and Laminated Iron," Trans. AIEE, Part I, Vol. 78, pp. 169-180 (1959).
7. P. Appun and H. Weh, "Wirbelströme in Feststehenden Teil von Zugmagneten zur Magnetischen Aufhängung von Fahrzeugen," ETZ, Vol. 91, pp. 623-627 (1971).
8. B.R. Cheo and E. Levi, "Modal Structures of the Electromagnetic Fields in a Moving Laminated Medium," to be submitted.

## II. SYSTEMS

### A. COMMUNICATIONS

### B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

### C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

### D. SYSTEMS, CONTROL AND NETWORKS

### E. DATA PROCESSING

## COMPARISON OF SEQUENTIAL PARTITION DETECTOR (SPD) WITH OTHER NONPARAMETRIC SEQUENTIAL DETECTORS

R. F. Dwyer and L. Kurz

In this report, the SPD will be compared with two sequential nonparametric detection techniques developed by Chadwick and Kurz.<sup>1</sup> They considered, essentially, a robust sequential sign test and its dual counterpart. The dual test uses two distinct nonparametric tests on the same received data sample. The transmitted wave form consists of either a signal,  $s_0$ , of level  $u_0$ , or a signal,  $s_1$ , of level  $u_1$ . The problem is to decide which signal was transmitted in the presence of additive noise. The dual formulation uses one of its tests to sequentially evaluate if  $s_0$  was transmitted, while the other test is processing the received data for the presence of  $s_1$ . If both tests agree, (i.e.,  $s_0$  is accepted and  $s_1$  rejected, or vice versa) to which signal was transmitted, a decision is made. However, if they are in disagreement, another sample is taken.

The classical sign test employs the statistic,

$$y_n = \frac{1}{n} \sum_{i=1}^n z_i$$

where

$$z_i = \begin{cases} 1 & \text{if } x_i \geq 0 \\ 0 & \text{if } x_i < 0 \end{cases}$$

and  $x_i$  ( $i = 1, 2, \dots, n$ ) are i.i.d. r.v. of the received data. The thresholds were derived from the Chebyshev and Chernoff bounds, respectively.

$$\frac{1}{2} - \frac{1}{2} \left( \frac{1}{1 + \alpha n} \right)^{1/2} < y_n < \frac{1}{2} + \frac{1}{2} \left( \frac{1}{1 + \alpha n} \right)^{1/2} \quad (1)$$

and

$$\frac{1}{2} - \frac{1}{2} \left( 1 - \alpha^{2/n} \right)^{1/2} < y_n < \frac{1}{2} + \frac{1}{2} \left( 1 - \alpha^{2/n} \right)^{1/2} \quad (2)$$

A simulation was conducted in order to compare the results of Ref. 1 for the sequential sign test with the SPD. The SNR was defined as  $\Delta_1/\sigma^2$ , where  $\Delta_1$  represents the shift in signal level and  $\sigma^2$  is the noise variance of  $N(0, \sigma)$ . Table I represents the results of 40 trials for two cases,  $\Delta_1 = .92$  and  $\Delta_1 = .46$  and the average sample size (ASN) is the average of the outcomes. Table II gives the results from Chadwick and Kurz<sup>1</sup> for the same  $\alpha$  and SNR used in Table I. Two points should be mentioned:

(1) The SPD has the advantage of being able to specify  $\alpha$ ,  $\beta$ , and adjusting ASN by varying  $\Delta_1$  in advance, whereas it is not usually possible to do this for the sequential sign test; and (2) by increasing  $m$ , the SPD represents a generalization of the sequential sign test.

TABLE I. ASN for the SPD under a shift of the mean alternative.

	$m = 2$	$m = 4$		$m = 2$	$m = 4$
$\Delta_1$	.46	.46		.92	.92
$\Delta$	.92	.92		.92	.92
$\beta(\Delta)$	.0005	.0005		.004	.004
ASN	24.6	19.95		13.95	13.875

$\alpha = .01$ ; SNR = 0 dB

TABLE II. ASN for the sign test.

	CHEBYSHEV THRESHOLD	CHERNOFF THRESHOLD
$\beta(\Delta)$	.004	.004
ASN	44.5	9.5

$\alpha = .01$ ; SNR = 0 dB

A closer look at the SPD for  $m = 2$  reveals an interesting similarity. Recall from Ref. 2 that the test statistic for the SPD is given by:

$$b < T_n = \sum_i^n \sum_k^2 b_k n_{ik} < a \quad (3)$$

or

$$b < b_1(n - \sum_i^n n_{i2}) + b_2 \sum_i^n n_{i2} < a$$

since,

$$n = \sum_i^n n_{i1} + \sum_i^n n_{i2}$$

Therefore,

$$\frac{b}{n(b_2 - b_1)} - \frac{b_1}{b_2 - b_1} < \frac{1}{n} \sum_i^n n_{i2} < \frac{a}{n(b_2 - b_1)} - \frac{b_1}{(b_2 - b_1)}$$

if  $b_2 = -b_1$  and  $b_1 = -|b_1|$ , then

$$\frac{b}{n(b_2 + b_1)} + \frac{1}{2} < \frac{1}{n} \sum_i^n n_{i2} < \frac{a}{n(b_2 + b_1)} + \frac{1}{2} \quad (4)$$

Notice that Eq. (4) has the same form as Equations (1) and (2). If  $n \rightarrow \infty$ , all 3 test thresholds approach one-half. However, for  $n$  small, the thresholds are quite different. In Eq. (4), the thresholds depend upon  $\alpha$ ,  $\beta$  and the scores which in turn are derived from the likelihood ratio. Therefore, all the information known about the sequential test is given in the thresholds. This is not the case, however, in Equations (1) and (2). Also, since  $a$  and  $b$  are derived from  $\alpha$  and  $\beta$ , no initial sample size must be specified for the SPD. On the other hand, in the sequential sign test, a minimum sample size must be given to insure a given  $\alpha$ . Table III compares the thresholds as a function of  $n$  for the Chebyshev, Chernoff and SPD bounds. Note that for  $n$  small, the SPD has larger thresholds which insure the specified error probabilities.

TABLE III. Thresholds for the sign test using Chebyshev, Chernoff and SPD bounds and  $\alpha = .0067$ .

n	CHEBYSHEV	CHERNOFF	SPD
1	$1.66 \times 10^{-3}$ ; .998	$1.12 \times 10^{-5}$ ; .1	-4.5; 5.5
2	$33 \times 10^{-3}$ ; .996	$1.67 \times 10^{-3}$ ; .998	-2; 3
5	$8 \times 10^{-3}$ ; .99	.035; .965	-.5; 1.5
10	.016; .98	.102; .898	0; 1
	Lower; Upper	Lower; Upper	Lower; Upper

### A. The Dual Test

The ASN can be reduced further by using the dual test formulation. Let one of two signals ( $s_0$  and  $s_1$ ) be transmitted as before. However, the hypothesis now will be the noise only condition. The presence of  $s_0$  and  $s_1$  will be tested separately against the hypothesis. This requires two tests as discussed above. Figure 1 represents the testing procedure for the dual test.

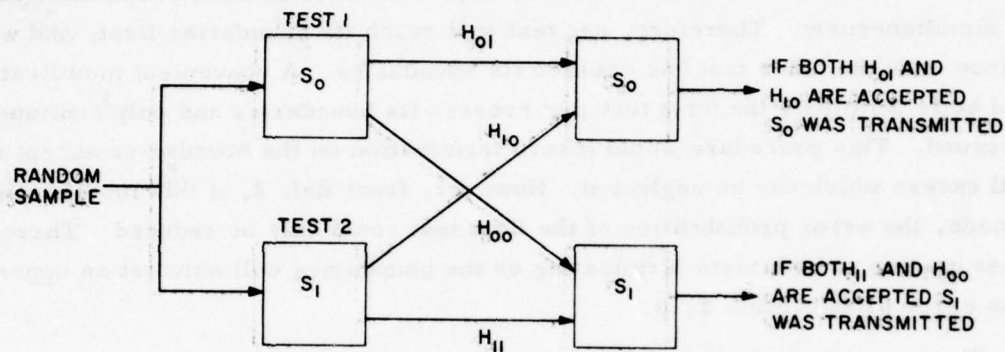


Fig. 1. Dual test sequential testing procedure.

Let,

$$H_{00}: F(x), \quad H_{01}: G(x) = F(x + s_0)$$

and

$$H_{10}: F(x), \quad H_{11}: G(x) = F(x - s_1)$$

Then, if both  $H_{01}$  and  $H_{10}$  are accepted,  $s_0$  was transmitted, or if both  $H_{11}$  and  $H_{00}$  are accepted,  $s_1$  was transmitted.

Since the tests are symmetrical, only the results obtained for sequentially detecting the presence of  $s_0$  will be given.

Consider the joint probability of accepting  $s_0$  and at the same time, rejecting the possibility of  $s_1$  being present when  $s_1$  was indeed transmitted

$$\begin{aligned} &P(T_{no} \geq T_a/s_1; T_{n1} \leq T_b/s_1) \\ &= P(T_{no} \geq T_a/s_1) P(T_{n1} \leq T_b/s_1) \\ &= P(T_{no} = T_a/s_1) P(T_{n1} = T_b/s_1) \\ &= \alpha_0 \beta_1 = a_D \end{aligned} \tag{5}$$

and

$$P(T_{no} \leq T_b/s_0; T_{n1} \geq T_a/s_0) = \beta_0 \alpha_1 = \beta_D \tag{6}$$

Where  $T_a, T_b$  indicate upper and lower thresholds,  $\alpha_D$  = false alarm probability for dual test,  $\beta_D$  = false dismissal probability for dual test, and  $\alpha_1, \alpha_0, \beta_1, \beta_0$  are the individual false alarm and false dismissal probabilities, respectively. The tests are considered independent and they terminate on the boundaries.

The assumption that both tests terminate on the boundaries needs some explanation. It is reasonable to assume that, usually, both tests will not reach their boundaries simultaneously. Therefore, one test will reach its boundaries first, and will continue until the other test has crossed its boundaries. A convenient modification would be to terminate the first test that crosses its boundaries and only continue with the second. This procedure would insure termination on the boundaries except for a small excess which can be neglected. However, from Ref. 2, if this modification was not made, the error probabilities of the first test could only be reduced. Therefore, the assumption of both tests terminating on the boundaries will only set an upper limit on the error probabilities  $\alpha, \beta$ .

The overall probability of false alarm,  $\alpha_D$ , and false dismissal,  $\beta_D$ , are a product of the individual  $\alpha$  and  $\beta$ . Therefore, given the desired  $\alpha_D$  and  $\beta_D$  for the dual test, the individual tests  $\alpha$  and  $\beta$  can be reduced proportionally, which is equivalent to lowering the thresholds.

#### Example

Given  $\alpha_D = .01$  and  $\beta_D = .004$ , then the individual  $\alpha$  and  $\beta$  become

$$\alpha_D = \alpha_0 \beta_1 = .01; \beta_D = \alpha_1 \beta_0 = .004$$

Let

$$\alpha_0 = \beta_1 \quad \text{and} \quad \alpha_1 = \beta_0$$

Therefore,

$$\alpha_0 = \beta_1 = .1; \alpha_1 = \beta_0 = .0632$$

Figure 2 gives the calculated results of using the above  $\alpha_0, \beta_0$  for  $m = 2, 4$ .

A simulation was conducted using the SPD in a dual test formulation in order to compare the predicted performance given in Figure 2. Table IV compares the results for 40 trials of a simulation with the predicted performance given in Figure 2. Note that the ASN obtained in the simulation corresponds very well with the ASN predicted in Fig. 2 for  $\Delta = .92$ .

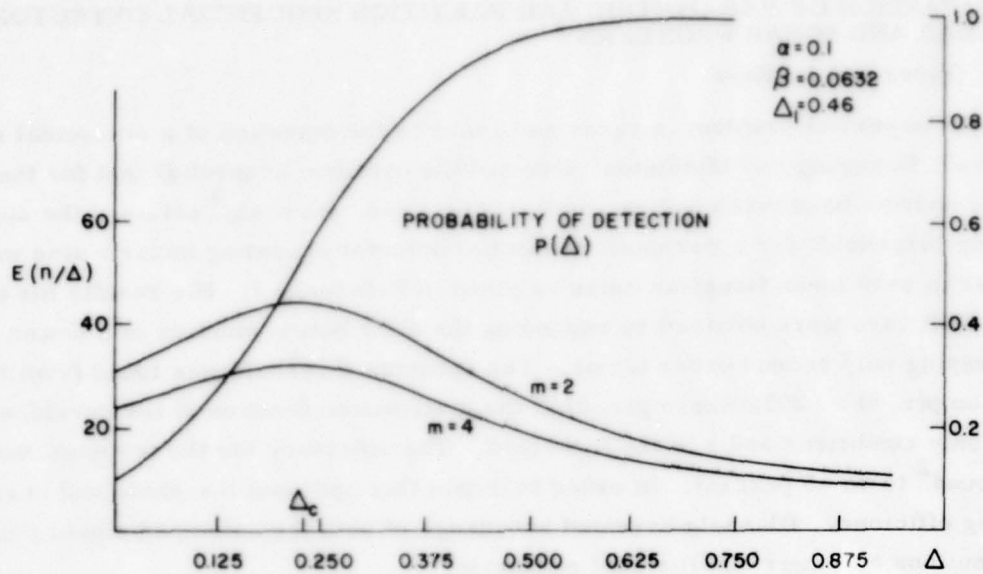


Fig. 2. ASN shift alternative.

TABLE IV. ASN for the SPD under a dual test formulation.

	SPD Predicted	Simulation
m = 2	13.05	15.7
m = 4	10.7	10.7

 $\Delta = .92$ 

Joint Services Technical Advisory Committee  
F44620-74-C-0056

R. F. Dwyer and L. Kurz

Naval Underwater Systems Center

## REFERENCES

1. M. Chadwick and L. Kurz, "Two Sequential Nonparametric Sequential Procedures," Information and Control Vol. 13, No. 5, pp. 403-428 (November 1968).
2. R. F. Dwyer and L. Kurz, "Sequential Partition Detectors," Journal of Cybernetics, No. 8, pp. 133-157 (1978).

# A COMPARISON OF PARAMETRIC AND PARTITION SEQUENTIAL DETECTORS IN RADAR AND SONAR PROBLEMS

R. F. Dwyer and L. Kurz

An important problem in radar and sonar is the detection of a sinusoidal signal in noise. Bussgang and Middleton<sup>1</sup> obtained the optimum sequential test for the envelope of narrow-band noise and an additive sine wave. Blasbalg<sup>2</sup> obtained the optimum "slicing threshold" for a Bernoulli sequential detector operating under a sine wave carrier in zero mean Gaussian noise as given in Reference 1. His results for the small SNR case were obtained by expanding the distribution function in a power series and keeping only second order terms. The optimum threshold was found from the equation  $p(r, 0) = .205$ , where  $p(r, 0)$  is the distribution function of the envelope under noise only conditions and  $r$  is the threshold. The efficiency for the optimum threshold was found<sup>2</sup> to be 65 percent. In order to obtain this optimum threshold and corresponding efficiency, Blasbalg assumed knowledge of both the noise and signal plus noise distributions and their small signal expansion.

In a recent paper,<sup>3</sup> the authors introduced the theory of sequential partition detectors (SPD). For the Lehmann alternative, which is a good model for radar and sonar problems, with  $m = 2$ , the optimum quantile yielded the efficiency of 65 percent. This result applies to all distributions under the Lehmann alternative satisfying the small SNR conditions. However, as demonstrated in Ref. 3, the small SNR condition can also be relaxed while maintaining good performance estimates. Also, for higher  $m$ , the SPD represents a generalization of the Bernoulli sequential detector.

Now, consider the problem of constructing a parametric sequential test for the incoherent detection of a sine wave in Gaussian,  $N(0, 1)$ , noise. This is the case discussed in Ref. 4 except that the output envelope will be squared before forming the test statistic. It is well known<sup>4</sup> that the quadratically detected output has a non-central chi-square distribution with p.d.f.

$$g(x) = \frac{1}{2\sigma^2} e^{-\frac{x+a_0^2}{2\sigma^2}} I_0(a_0\sqrt{x}/\sigma^2), \quad x \geq 0 \quad (1)$$

$$= 0, \quad x < 0$$

Where  $I_0(\cdot)$  is the modified Bessel function of first kind and zero order,  $\sigma^2$  represents the noise variance;  $x$  is the squared output of a narrowband filter, and  $a_0$  is the amplitude of the sinusoidal signal. For  $a_0 = 0$ , Eq. (1) reduces to the Rayleigh distribution with p.d.f.

$$f(x) = \frac{1}{2\sigma^2} e^{-x/2\sigma^2}, \quad x \geq 0 \quad (2)$$

$$= 0, \quad x < 0$$

Also, the signal-to-noise ratio, SNR is defined as,  $SNR = S_1 = a_0^2/2\sigma^2$ .

Then, using Wald's loglikelihood ratio, the sequential test statistic is expressed as:

$$S_n = \sum_{i=1}^n \lambda_i \quad (3)$$

where

$$\lambda_i = -a_0^2/2\sigma^2 + \ln [I_0(a_0/\sigma^2 \sqrt{x})] \quad (4)$$

The thresholds for the square-law sequential detector are,

$$b < S_n < a$$

and the test terminates with acceptance of  $H_0$  or  $H_1$  if  $S_n \leq b$  or  $S_n \geq a$ , respectively. Notice that the thresholds  $(a, b)$  for the parametric detector are identical to the ones used in the SPD.<sup>3</sup>

For small SNR, Eq. (4) reduces to

$$\lambda_i = -a_0^2/2\sigma^2 (1 + \frac{1}{2} a_0^2/2\sigma^2) + a_0^2/2\sigma^2 x/2\sigma^2 \quad (5)$$

and

$$E(\lambda_i/S) = SS_1 - \frac{1}{2} S_1^2$$

Where  $S$  and  $S_1$  play the same role of the true and selected alternatives as  $\delta$  and  $\delta_1$  do in the SPD.

Since  $S_1, S$  are assumed small, the value of  $t = t_0(S) = 0$  which solves the moment generating function,  $\phi(t) = 1$ , can be given in terms of the mean and variance of  $\lambda_i$ . This follows directly from the results in Ref. 3.

Then,

$$t_0(S) = -2 \frac{E(\lambda_i)}{\sigma_{\lambda_i}^2} = 1 - 2S/S_1 \quad (6)$$

Recall<sup>3</sup> that for the SPD,  $t_0(\delta) = 1 - 2\delta/\delta_1$ ,

and

$$E(T_i/\delta) = (\delta \delta_1 - \frac{1}{2} \delta_1^2) \sum_k^m A_k^2 / P_{ok}.$$

If the relationship between  $\delta$  and  $S$  was known, an efficiency expression could be obtained in a simple form, comparing the optimum parametric sequential test with the SPD. To that end, a technique which measures the separation between the hypothesis and alternative will be considered.

The change in mean is often used as a performance measuring parameter in parametric tests.

Let,  $H_0: F(x)$  and  $H_1: G(x)$ . Then,

$$E(x/H_1) = \int_{-\infty}^{\infty} x g(x) dx, \quad \sigma_{x/H_1}^2 = E[(x - E(x/H))^2]$$

The change in the mean is expressed as,

$$\Delta m = \int_{-\infty}^{\infty} x(g(x) - f(x)) dx \quad (7)$$

Using Eqs. (1) and (2), it can be shown that

$$\Delta m \approx a_0^2 \quad (8)$$

Therefore, for small  $a_0$ , the change in the mean is given by the square of the sinusoidal amplitude.

For the SPD under a Lehmann alternative, i.e.,

$$H_0: F(x) \quad \text{and} \quad H_1: G(x) = F^{1+\delta}(x)$$

The change in the mean, using Eq. (2) as the reference p.d.f. is given by,

$$\Delta m = \int_0^{\infty} x(1+\delta) F^{\delta}(x) f(x) dx - \int_0^{\infty} x f(x) dx.$$

Where

$$F(x) = (1 - e^{-x/2\sigma^2})$$

Let

$$y = e^{-x/2\sigma^2}$$

Then,

$$\Delta m = -(1+\delta)2\sigma^2 \int_0^1 \ln(1-y) y dy - 2\sigma^2 \quad (9)$$

For  $\delta \ll 1$ , Eq. (7) reduces to

$$\Delta m = 2\sigma^2 \delta \quad (10)$$

From Eqs. (8) and (10), the Lehmann indexing parameter  $\delta$  is equal to the SNR of the parametric sequential test, i.e.,

$$\delta = a_o^2/2\sigma^2$$

Therefore, the efficiency expression is given by,

$$\mathcal{E}(S_p/T_{Lm}) = \sum_k^m A_k^2/P_{ok} \quad (11)$$

$$\delta, a_o \rightarrow 0$$

where  $S_p$  represents the parametric sequential detector and the notation of Ref. 3 has been used. Typical values for Eq. (11) can be obtained from Table I.

TABLE I. Efficiency for SPD under shift and Lehmann alternatives

m	Shift Alternative Gaussian Noise, $N(0, 1)$	Lehmann Alternative
2	.636	.65
4	.88	.89
6	.94	.95
8	.965	.968
10	.977	.977

Unfortunately, Eq. (11) only applies for small  $\delta$ . For other values, Eq. (9) must be solved as a function of  $\delta$ , which is difficult in general. Gotz and Kurz<sup>5</sup> have evaluated by numerical methods, the change in the mean and also the change in variance

as a function of the Lehmann alternative using a reference c.d.f. of  $N(0, 1)$ .

A simulation was conducted comparing the sequential square-law (optimum) detector with the SPD. The input consisted of additive zero mean Gaussian noise and a constant amplitude sinusoidal signal which was passed through a narrowband filter and squared. The quantiles for the SPD were estimated from a training sample at the output of the square law device for  $m = 2, 4$ . Figure 1 shows the results for 20 trials of

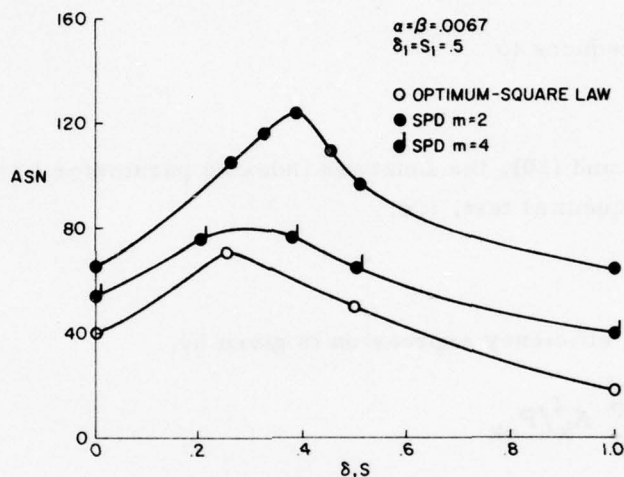


Fig. 1. ASN - Optimum vs. SPD (Lehmann alternative) - simulation.

the simulation. Both the selected Lehmann,  $\delta_1$ , and parametric,  $S_1$ , parameters were set to one-half. As seen from Fig. 1, the ASN for the SPD approaches the optimum detector's performance very rapidly as  $m$  increases.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

R. F. Dwyer and L. Kurz

Naval Underwater Systems Center

#### REFERENCES

1. J. Busgang and D. Middleton, "Optimum Sequential Detection of Signals in Noise," IRE Trans. on Info. Theory, Vol. IT-1 (December 1955).
2. H. Blasbalg, "The Relationship of Sequential Filter Theory to Information Theory and its Application to the Detection of Signals in Noise by Bernoulli Trials," IRE Trans. on Info. Theory (June 1957).
3. R. F. Dwyer and L. Kurz, "Sequential Partition Detectors," Journal of Cybernetics, No. 8, pp. 133-157 (1978).
4. C. Halstrom, "Statistical Theory of Signal Detection Pergamon (1968).
5. S. B. Gotz and L. Kurz, "The Class of Lehmann Alternatives as Means for Evaluation of Performance in Robust Detection," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. 452.41-76, pp. 242-248 (1976).

## M. Kavehrad and L. Kurz

### A. Design of Recursive Equalizers

Fig. 1. Data transmission system.

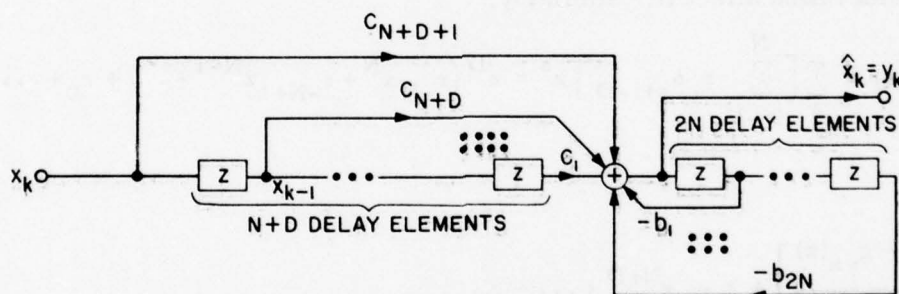


Fig. 2. Recursive equalizer structure.

$$\Gamma(z) = [\rho^2 B(z)]^{-1} [g_{yx}(z)/B(z^{-1})]_+ \quad (1)$$

where  $g_{vx}(z)$  is the cross-spectral density function and  $\rho^2$  and  $B(z)$  are taken from the

canonical factorization of spectral density function

$$g_{xx}(z) = \rho^2 B(z) B(z^{-1}) \quad (2)$$

The symbol (+) in Eq. (1) indicates that only the non-negative powers of  $z$  are retained. The  $g_{xx}(z)$  in Eq. (2) is obtained from

$$g_{xx}(z) = \sum_s \left( \sum_{i=-N}^N \sum_{j=-N}^N r_i r_j \delta_{s+j-1} + \sigma_n^2 \delta_s \right) z^s \quad (3)$$

which is of the form

$$g_{xx}(z) = d_{-2N} z^{-2N} + d_{-2N+1} z^{-2N+1} + \dots + d_{-1} z^{-1} + d_0 + d_1 z + \dots + d_{2N} z^{2N} \quad (4)$$

with

$$d_{2N-i} = \sum_{j=0}^i r_{-N+j} r_{N-i+j} + \sigma_n^2 \delta_{2N-i} \quad (5)$$

It should be noted that in deriving Eq. (4) it is assumed that only intersymbol interferences which occurs within the  $N$  preceding and succeeding symbols of the symbol under consideration affect it. Similarly,

$$g_{yx}(z) = \sum_s \left[ \sum_{i=-N}^N r_i \delta_{s+i-D} \right] z^s = z^D (r_{-N} z^N + r_{-N+1} z^{N-1} + \dots + r_0 + \dots + r_N z^{-N}) \quad (6)$$

and

$$\left[ \frac{\frac{1}{2} g_{yx}(z)}{\rho B(z^{-1})} \right]_+ = c_1 z^{N+D} + \dots + c_{N+D+1} \quad (7)$$

while

$$B(z) = 1 + b_1 z + \dots + b_{2N} z^{2N} \quad (8)$$

The values of  $b$ 's and  $c$ 's give us the design parameters for the recursive equalizer of Figure 2.

To obtain specific comparison results with the equalizer of Ref. 1, we assume that the timing jitter immune partial response signals are of minimum bandwidth

(twice the Nyquist bandwidth) and are of the Kretzmer form.<sup>3</sup> As was shown in Ref. 1, the overall spectrum for timing-jitter-immune partial response signals is

$$R(\omega) = T \left( 1 - \frac{T}{2\pi} |\omega| \right) \sum_{k=-N}^N r_k e^{-j\omega k T} \quad (9)$$

Using the design equations, the specific results for the taps of Kretzmer signals are:

- I.  $r_0 = r_1 = r_{-1} = 1$   
 $c_1 = .346 \quad c_2 = .168 \quad c_3 = .14$   
 $b_1 = .514 \quad b_2 = .345$
- II.  $r_0 = r_2 = r_{-2} = 1 \quad r_1 = r_{-1} = 2$   
 $c_1 = .172 \quad c_2 = .243 \quad c_3 = .071 \quad c_4 = .142$   
 $b_1 = b_2 = b_3 = .59$
- III.  $r_0 = 1 \quad r_1 = r_{-1} = \frac{1}{2} \quad r_2 = r_{-2} = \frac{1}{2}$   
 $c_1 = -.19 \quad c_2 = .23 \quad c_3 = .29 \quad c_4 = .14 \quad c_5 = .089$   
 $b_1 = .212 \quad b_2 = -.22 \quad b_3 = -.21 \quad b_4 = .095$
- IV.  $r_0 = 1 \quad r_1 = r_{-1} = 0 \quad r_2 = r_{-2} = -1$   
 $c_1 = -.35 \quad c_2 = c_4 = 0 \quad c_3 = -.52 \quad c_5 = -.49$   
 $b_1 = b_3 = 0 \quad b_2 = .514 \quad b_4 = .35$
- V.  $r_1 = r_{-1} = r_3 + r_{-3} = 0 \quad r_2 = r_{-2} = 2 \quad r_0 = r_4 = r_{-4} = -1$   
 $c_1 = -.17 \quad c_2 = c_4 = c_6 = c_8 = 0 \quad c_3 = -.21 \quad c_5 = -.26 \quad c_7 = -.061$   
 $c_9 = .21$   
 $b_1 = b_3 = b_5 = b_7 = 0 \quad b_2 = -.325 \quad b_4 = 3.48 \quad b_6 = -1.69 \quad b_8 = .172$

There are two basic advantages of recursive equalization described above: the equalizer may be easily operated in an adaptive mode by adjusting the taps based on the estimation of channel conditions. it improves equalization in some cases up to 50%, basing the comparison on the signal-to-noise ratio degradation. The specific numerical comparisons for the Kretzmer signals are given in Table I.

TABLE I

Partial Response Signal	Signal-to-Noise Ratio Degradation in dB	
	Nonrecursive	Recursive
1 (ideal)	0	0
$D_L^{-1} + 1 + D_L$	1.80	1.90
$D_L^{-2} + 2 D_L^{-1} + 1 + 2 D_L + D_L^2$	19.1	19.5
$-\frac{1}{2} D_L^{-2} + \frac{1}{2} D_L^{-1} + 1$ $+ \frac{1}{2} D_L - \frac{1}{2} D_L^2$	2.10	2.40
$- D_L^{-2} + 1 - D_L^2$	3.90	3.70
$- D_L^{-4} + 2 D_L^{-2} - 1$ $+ 2 D_L^2 - D_L^4$	12.70	7.60
$D_L^{-9} - D_L^{-4} + 2 D_L^{-2} - 1$ $1 - 2 D_L^2 - D_L^4 + D_L^9$	13.6	6.9

$D_L$  - unit delay operator.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

M. Kavehrad and L. Kurz

#### REFERENCES

1. M. Kavehrad and L. Kurz, "Suppression of Timing Jitter in Partial Response Systems," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-7, pp. 291-295 (1977).
2. S. M. Fitch and L. Kurz, "Recursive Equalization in Data Transmission - A Design Procedure and Performance Evaluation,"

## ROBUST SCORE ESTIMATION FOR SIMPLE LINEAR RANK DETECTORS

I. M. Habib and L. Kurz

Nonparametric methods, which have found broad areas of applications in detection and estimation,<sup>1</sup> are based on the branch of statistical inference for which the parameter space associated with the underlying distributions cannot be represented by a finite set of real numbers. A commonly used property of nonparametric tests is that the level of the test is fixed irrespective of the distribution of the observables. For a given level, the rank tests represent a class of powerful tests. To preserve the power of the rank tests, the knowledge of scores is essential. If the scores of the rank tests are selected beforehand, based on some general considerations irrespective of the underlying distribution under the alternative, the test may have low power, especially if the selected scores are sufficiently different from those generated by the true distribution.<sup>2</sup> It is the purpose of this report to develop a robust (insensitive to changes in underlying distribution) procedure which generates proper scores based on the sample of observables for simple rank tests. A similar philosophy was used elsewhere<sup>3</sup> by the authors to develop score estimators for generalized quantile detectors.

## A. Score Estimation

Let the observables form a sample (vector)  $\underline{x}$  of size  $N$  and  $\ell_i(\underline{x})$  denote the value of the  $i$ -th smallest coordinate in  $\underline{x}$ . Denoting  $x^{(i)} = \ell_i(\underline{x})$ , we have

$$x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(N)} \quad (1)$$

where  $x^{(i)}$  is the  $i$ -th order statistic of  $\underline{x}$ . Let  $r_i(\underline{x})$  denote the number of  $x$ 's  $\leq x_i$ , i.e., the rank of  $x_i$  in Equation (1).

The statistic  $R_i = r_i(\underline{x})$  is called the rank of  $x_i$  and  $\underline{R}$  is the vector of the ranks of  $\underline{x}$ . The ranks are well defined only if the probability of coincidence of any pair of coordinates equals zero.

A simple linear rank test is given by

$$S = \sum_{i=1}^N c_i a(R_i) \quad (2)$$

where  $c_i$  is 0 or 1 and  $a(R_i)$  are the scores of the test. These scores are functions of c.d.f.,  $F$ , and p.d.f.,  $f$ , under the alternative.

The scores are generated by a score generating function  $\varpi(t)$ ,  $0 < t < 1$ , as follows:

$$a(i, f) = E[\varphi(u_N^{(i)}, f)] \quad 1 \leq i \leq N \quad (3)$$

where  $u^{(i)}$  is an ordered sample from the  $(0, 1)$  uniform distribution. Generally, the  $\varphi$  function form depends on the test parameter, i.e., a shift parameter or a scale parameter. For the shift parameter

$$\varphi(u, f) = \frac{f'[F^{-1}(u)]}{f[F^{-1}(u)]} \quad 0 \leq u \leq 1 \quad (4)$$

where  $f'$  is the derivative of  $f$  and  $F^{-1}(u)$  is the  $u$ -th quantile. For the scale parameter

$$\varphi_1(u, f) = -1 - F^{-1}(u) \frac{f'[F^{-1}(u)]}{f[F^{-1}(u)]} \quad 0 \leq u \leq 1 \quad (5)$$

If  $N$  is large and  $\varphi(u, f)$  is smooth in the neighborhood of  $i/N + 1$ , we have

$$a(i, f) = \varphi\left(\frac{i}{N+1}, f\right) \quad (6)$$

To generate  $\varphi(\cdot)$  and  $a(i, f)$  from  $x_i$ , we use an estimator shown in Figure 1. This procedure for estimating the scores fits well in the framework of nonparametric tests and improves the power of the tests under varying noise conditions and in the presence of outliers. The essential building blocks for the estimator of Fig. 1 were developed in References 4 and 5.

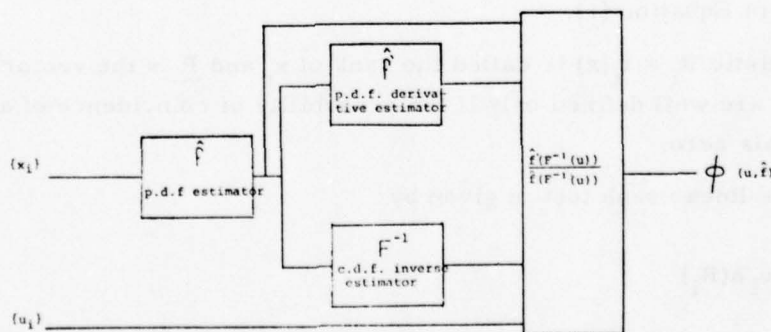


Fig. 1. Score estimator for linear rank tests.

B. An Example

Since the performance of each block of Fig. 1 was investigated in Refs. 4 and 5 yielding good to excellent results, this tends to indicate that the score estimation procedure, a simulation program was developed for the Van-der-Waerden test<sup>2</sup> and compared to the theoretical scores given by

$$a_i = \Phi^{-1}\left(\frac{i}{m+1}\right) \quad 1 \leq i \leq m$$

where

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^x \exp\left(-\frac{1}{2}y^2\right) dy$$

The results of the simulation study for  $m = 20$  are given in Table I.

TABLE I. Scores of the Van-der-Waerden test.

Exact Scores	Estimated Scores
-1.67	-1.6
-1.31	-1.28
-1.07	-1.1
-0.88	-.9
-0.71	-.77
-0.57	-.6
-0.43	-.44
-.3	-.28
-.18	-.2
-0.06	-.04
0.06	.04
0.18	.28
0.3	0.34
0.43	.49
0.57	.58
.71	.7
.88	.75
1.07	.9
1.31	1.1
1.67	1.64

### C. Comparison With Other Methods

Hajek<sup>6</sup> considers a class of score estimation procedures for simple rank tests. In his notation, Hajek proceeds to generate scores as follows.

For  $N \geq 1$ , consider the simple linear rank test

$$S_n = \sum_{i=1}^n c_{ni} a_n(R_{ni})$$

where the ranks  $R_{n1}, \dots, R_{nn}$  are derived from a sequence of observations  $Y_{n1}, \dots, Y_{nn}$  and  $a_n$  are the scores.  $S_n$  may be written as

$$S_n(Y_{n1}, \dots, Y_{nn}) = \sum_{i=1}^n c_{ni} a_n \left( \sum_{j=1}^n u(Y_{ni} - Y_{nj}) \right)$$

where  $u(\cdot)$  is the unit step function.

Consider the increment of  $S_n$  if the observables  $Y_{ni}$ 's are shifted by a random multiple of some numbers  $d_{ni}$ , i.e., consider the statistic

$$S_n^* = S_n(Y_{n1} + \Delta_n d_{n1}, \dots, Y_{nn} + \Delta_n d_{nn})$$

Let

$$\begin{aligned} d_{ni} &= 1/\sqrt{n} & (1 \leq i \leq [\frac{n}{2}]) \\ &= 0 & (\frac{n}{2} \leq i < n) \end{aligned}$$

and let  $\Delta_n = \Delta_n(Y_1, \dots, Y_n)$  be any statistic such that

$$\Delta_n(\lambda Y_1 + u, \dots, \lambda Y_n + u) = \lambda \Delta_n(Y_1, \dots, Y_n) - \infty < u < \infty, \quad \lambda > 0$$

for which

$$P \lim_{f=f_j} \Delta_n(x_1 \dots x_n) = b_j < \infty \quad (1 \leq j \leq k)$$

where  $P$  denotes probability and  $f_j$  denotes the  $j$ -th p.d.f. Hajek suggest the following selections:

$$\Delta_n = E \left[ x_n^{(n+1-j_n)} - x_n^{(j_n)} \right] \quad \left( \frac{j_n}{n} \rightarrow \alpha, \quad 0 < \alpha \leq \frac{1}{2} \right)$$

$$b_j = E \left[ F_j^{-1}(1-\alpha) - F_j^{-1}(\alpha) \right]$$

$$S_{nj} = S_{nj}(x_1, \dots, x_n), \quad 1 \leq j \leq k$$

$$S_{nj}^* = S_{nj}(x_1 + \Delta_n d_{n1}, \dots, x_n + \Delta_n d_{nn})$$

The ratios

$$l_{nj} = \frac{S_{nj}^* - S_{nj}}{\sqrt{\text{var} S_{nj}}}$$

are then used for score estimation.

It is obvious that the method suggested by Hajek is more complicated than the method outlined in Section A. The processing time is increased significantly because of the ranking required and the choice of the  $\Delta$  statistic is not obvious for a given application. In addition, the procedure suggested by Hajek is not robust, e.g., if the samples are lightly contaminated the score will be taken as those for the pure distribution reducing the power of the test. If the contamination of the samples is severe, the computation of  $l_{nj}$  is different for the two underlying distributions. It is not clear how to modify Hajek's procedure and the approach fails completely.

The procedure developed here is easy and inexpensive to implement, does not require ranking and storing of samples and preserves the power of the test even if the samples are contaminated.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

I. M. Habib and L. Kurz

#### REFERENCES

1. Nonparametric Methods in Communications, P. Papantoni-Kazakos and D. Kazakos, editors, Marcel Dekker (1977).
2. J. Hajek and A. Sidak, Theory of Rank Tests, Academic Press (1967).
3. I. M. Habib and L. Kurz, "Score Estimation for the Generalized Quantile Detector," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, No. R-452.42-77, pp. 345-352 (1977).
4. I. M. Habib and L. Kurz, "Nonparametric Probability Density Estimation," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. 452.41-76, pp. 586-591 (1976).
5. I. M. Habib and L. Kurz, "On Channel Monitoring Techniques in the Presence of Impulsive Noise or Fading Effects," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77, pp. 275-290 (1977).
6. J. Hajek, "Miscellaneous Problems of Rank Test Theory in Nonparametric Techniques in Statistical Inference, M. L. Puri, editor, Cambridge Univ. Press (1970).

# EXTRAPOLATION OF BAND-LIMITED SIGNALS USING A STOCHASTIC APPROXIMATION ALGORITHM

I. Kadar and L. Kurz

Recently there has been renewed interest in extrapolating band-limited signals and computing their Fourier transform from noisy measurements which can only be taken over a finite segment of the signal.<sup>1-5</sup> This problem is of great importance in spectral analysis<sup>2-4,6,7</sup> in optical or antenna systems where one desires to extend the resolution of the instrument beyond the diffraction limit by super-resolution<sup>4,6,7</sup> or in image processing where the object (or the entire image) is reconstructed by extrapolating a truncated observed portion.<sup>4</sup> As was indicated by the authors previously, the approaches based on the prolate spheroidal wave functions<sup>2</sup> and the use of appropriate projection in the Hilbert space<sup>3</sup> lead to improperly posed problems. In the absence of noise, these algorithms become unstable in the limit. This problem can be circumvented by the use of appropriate stopping rules. On the other hand, in the presence of noise large errors occur for finite number of iterations and in most practical situations the algorithms will diverge. In a recent paper,<sup>5</sup> the authors introduced a stable algorithm, based on the Gladyshev theorem, which guarantees the convergence of the extrapolation or restoration algorithm with probability one and in the mean-square sense even in the presence of corrupting severe noise outliers. Another version of the stochastic approximation algorithm is presented here. The approach uses expansions into prolate spheroidal wave functions as in Ref. 2, yet guarantees convergence with probability one and in the mean-square sense in the presence of noise. The new approach has the potential of being generalized to two-dimensions, including the use of other orthogonal function sets, thus opening the methodology to a broad class of problems in image reconstruction.

## A. An Algorithm Based on Stochastic Approximation

Using the approach developed in Ref. 2 and applying a vector Robbins-Monro stochastic approximation (RMSA)<sup>8</sup> algorithm directly to estimate the coefficients in the prolate spheroidal expansion method in a manner similar to the one used for estimation of density functions,<sup>9</sup> one arrives at the new algorithm.

The prolate spheroidal expansion method in Ref. 2 yields for  $f(t)$  in terms of  $g(t)$

$$f(t) = \sum_{k=1}^{\infty} g_k \sqrt{\lambda_k} \phi_k(t)$$

where

$$g_k = \left(1/\lambda_k\right) \int_{-\infty}^{\infty} g(t) \phi_k(t) dt$$

$$\lim_{N \rightarrow \infty} \int_{-T}^T \left[ g(t) - \sum_{k=1}^N g_k \frac{\phi_k(t)}{\sqrt{\lambda_k}} \right]^2 dt \rightarrow 0$$

and  $\{\phi_k(t)\}$  are the prolate spheroidal wave functions generated by

$$\int_{-T}^T \phi_k(\tau) \frac{\sin \sigma(t-\tau)}{\pi(t-\tau)} d\tau = \lambda_k \phi_k(t) .$$

Now, if  $g(t)$  is corrupted by noise, i.e.,

$$y(t) = g(t) + v(t)$$

one can consider the problem as if  $g(t)$  is unknown (or cannot be directly observed) and we estimate an approximation to  $y(t)$  using a sequence of independent samples  $t_k$ ,  $k=1, 2, \dots$ . A block sample of  $y(t)$  is approximated as

$$\underline{y}(t) = \sum_{k=1}^n \underline{a}_k^* \underline{\psi}_k(t) \equiv \underline{a}^{*T} \underline{\psi}(t)$$

where  $\underline{\psi}_k(t) \equiv \phi_k(t)/\sqrt{\lambda_k}$ ,  $\underline{a}^*$  is an  $n$ -vector of unknown coefficients. The problem is to find  $\underline{a}^*$  which minimizes the integral-square-error (ISE) criterion which is given by

$$I(\underline{a}) = \int_{\Omega} (\underline{y}(t) - \underline{a}^T \underline{\psi}(t))^2 dt$$

where

$$dt \equiv \sum_{k=1}^m dt_k, \quad \Omega = (-T, T)$$

The value of  $\underline{a} = \underline{a}^*$  which minimizes this criterion is<sup>10</sup>

$$\underline{a}^* = \left[ \int_{\Omega} \underline{\psi}(t) \underline{\psi}^T(t) dt \right]^{-1} \int_{\Omega} \underline{\psi}(t) \underline{y}(t) dt$$

To evaluate  $\underline{a}^*$  numerically, using the RMSA, one forms

$$\underline{a}_{k+1} = \underline{a}_k - (A/k) \underline{Y}(\underline{a}_k, t_k)$$

where  $A$  is a diagonal matrix, the function  $\underline{Y}(\underline{a}, t)$  is chosen so that

$$E[\underline{Y}(\underline{a}, t)/\underline{a}] = \frac{\partial I}{\partial \underline{a}} = - \int_{\Omega} \underline{\psi}(\eta) \underline{g}(\eta) d\eta + \left[ \int_{\Omega} \underline{\psi}(\eta) \underline{\psi}^T(\eta) d\eta \right] \underline{a}$$

$$\eta = \text{col}(\eta_1, \dots, \eta_m), \quad d\eta = \prod_{i=1}^m d\eta_i$$

The function  $\underline{Y}(\underline{a}, t)$  satisfying the above condition is given by

$$\underline{Y}(\underline{a}, t) = -\underline{\beta}(t) + K \underline{a}$$

where

$$\underline{\beta}(t) = \begin{cases} \underline{\psi}(t) & \text{if } t \in \Omega \\ 0 & \text{if } t \notin \Omega \end{cases}$$

and

$$K = \int_{\Omega} \underline{\psi}(t) \underline{\psi}^T(t) dt$$

which can be evaluated directly from the knowledge of the eigenfunctions. The vector RMSA algorithm can be written as

$$\underline{a}_{k+1} = \underline{a}_k + (A/k) (\underline{\beta}(t_k) - K \underline{a}_k)$$

which converges in the mean-square sense and w.p. 1.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

I. Kadar and L. Kurz

Grumman Aerospace Corp.

#### REFERENCES

1. G. A. Viano, "On the Extrapolation of Optical Image Data," *Journal of Mathematical Physics*, Vol. 17, No. 7 (July 1976).
2. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-Limited Extrapolation," *IEEE Trans. on Circuits and Systems*, Vol. CAS-22 No. 9 (September 1975).
3. D. C. Youla, "Generalized Image Restoration by the Method of Alternating Orthogonal Projections," *Progress Report No. 41 to JSTAC, Polytech. Inst. of New York*, Report No. R-452.41-76, pp. 544-559 (1976).
4. R. W. Gerchberg, "Super-Resolution through Error Energy Reduction," *Optica Actor*, Vol. 21, No. 9 (1974).
5. I. Kadar and L. Kurz, "A Robustized Vector Recursive Stabilizer Algorithm for Image Restoration," *Progress Report No. 42 to JSTAC, Polytech. Inst. of New York*, Report No. R-452.42-77, pp. 536-546 (1977).
6. H. C. Andrews, *Computer Techniques in Image Processing*, Academic Press (1970).

7. A. Papoulis, Systems and Transforms with Applications in Optics, McGraw-Hill (1968).
8. H. Robbins and S. Monro, "A Stochastic Approximation Method," Ann. Math. Stat., Vol. 22 (1951).
9. K. L. Kashyap and C. C. Blaydon, "Estimation of Probability Density and Distribution Functions," IEEE Trans. on Info. Theory, Vol. IT-14, No. 4 (July 1968).

## INFLUENCE OF NONLINEARITIES ON THE PERFORMANCE OF PARTIAL RESPONSE SYSTEMS

M. Kavehrad and L. Kurz

In this report, the influence of nonlinearities on PAM systems using partial response techniques will be investigated. The report represents an extension and generalization of previous results.<sup>1,2</sup>

Such an investigation permits a meaningful study of practical systems because even if nonlinearity is not added to the system to suppress the effect of impulsive noise, all real systems have a finite dynamic range of linear operation and act as a soft limiter with a high saturation level.

In particular, the effect of nonlinearities on intersymbol interference is considered. The generalization of the equalizer design in the frequency domain is presented which includes the constraints of suppression of performance degradation due to nonlinearities. Several partial response signals which are of practical interest are considered and, in each case, the receiving filter (equalizer) characteristics assuming Square Root of Raised-Cosine as the system input, are obtained. Finally, a comparison between a standard pulse-amplitude-modulation (PAM) system and a partial response system in the presence of nonlinearities is given.

A. Suppression of Impulsive Noise by Addition of Nonlinearity in Partial Response Systems

The problem of suppressing the large excursions of the impulsive noise by addition of a preselected nonlinearity (naturally there is some nonlinearity in any system over which one has no control) is of great interest in the study of data transmission systems.

Impulsive noise is usually characterized by high amplitude bursts in relatively short intervals and the distribution of the amplitudes have heavier tails than can be represented by a Gaussian distribution. During these intervals the noise is assumed to be a sample function of a random process with a heavy-tailed distribution of various types (Cauchy, generalized exponential, etc.).

Although impulsive noise is present for only short intervals of time, a major part of the error rate in the system may be caused by it.

For example, in telephone networks, most of the errors in relatively low speed transmission systems is caused by impulsive noise which usually arises from electrical contacts that cause a transient effect in the switching and signaling equipment.

One way to suppress impulsive noise is to clip or limit the received waveform

by addition of nonlinearity before the receiving filter. Different types of nonlinearities have been used for this purpose.

A useful representation of a large class of such nonlinearities is the third power polynomial device expressed by  $y = a_1 x + a_2 x^3$  shown in Fig. 1, where by proper selection of the constants  $a_1, a_2$ , the curve resembles the optimal nonlinearity found by Rappaport and Kurz.<sup>3</sup>

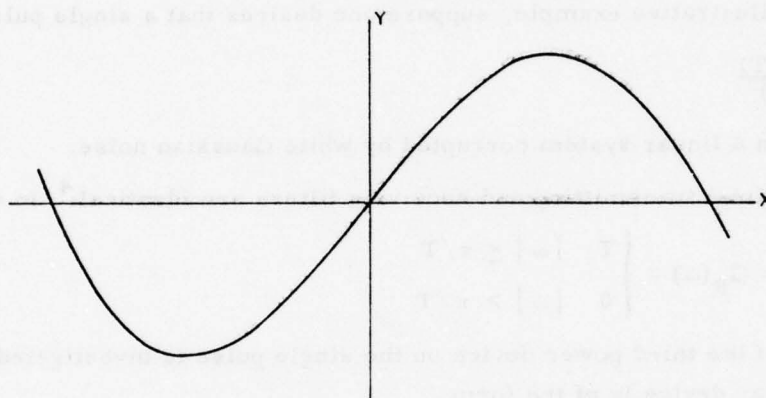


Fig. 1. Third power nonlinearity.

It should be observed that the shape of the nonlinearity is not of critical importance; the main problem is just to suppress the large excursions of burst noise.

### B. System Description

The transmission system model considered in this report is illustrated in Figure 2. Pulse Amplitude Modulated (PAM) signals are passed through a channel  $\tilde{C}(\omega)$ . It

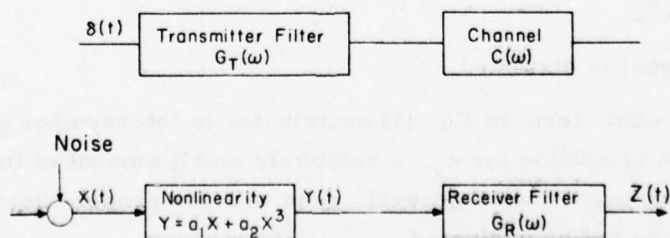


Fig. 2. Transmitting and receiving system model with intersymbol interference.

will be assumed that the channel has an ideal low-pass characteristic; even if it does not, we may still lump the more general response together with the transmitter.

Furthermore, it is not necessary to be restricted to baseband transmission. The effective noise is assumed to appear at the input to the receiver, which consists of a memoryless nonlinearity followed by a linear filter (equalizer),  $G_R(\omega)$ . In a linear system, intersymbol interference may be considered independent of addition noise, but if the received signal is passed through a memoryless nonlinearity, distortion will take place, i.e., a change in the expected value of the received waveform. This distortion is a function of the transmitted signal, the noise and the nonlinearity.

As an illustrative example, suppose one desires that a single pulse of the form

$$\frac{\sin(\pi t/T)}{(\pi t/T)}$$

be received in a linear system corrupted by white Gaussian noise.

The optimal transmitting and receiving filters are identical.<sup>4</sup> In this case

$$G_T(\omega) = G_R(\omega) = \begin{cases} T & |\omega| \leq \pi/T \\ 0 & |\omega| > \pi/T \end{cases}$$

If the effect of the third power device on the single pulse is investigated, the input signal to nonlinear device is of the form

$$x(t) = \frac{\sin(\pi t/T)}{(\pi t/T)} + \eta(t)$$

where  $\eta(t)$  is Gaussian noise of zero mean and  $\sigma_x^2$  variance, the expected value of the output waveform is

$$E\{y(t)\} = A \frac{\sin(\pi t/T)}{(\pi t/T)} + a_2 \left[ \frac{\sin(\pi t/T)}{(\pi t/T)} \right]^3 \quad (1)$$

where

$$A = a_1 + 3\sigma_x^2 a_2 \quad (2)$$

so the received signal is distorted.

Clearly, the cubic term in Eq. (1) contributes to intersymbol interference. Thus, by proper selection of a value for  $a_2$ , a relatively small amount of intersymbol interference can be introduced, but in general, even with the proper choice of  $a_2$ , the nonlinear distortion may not be neglected.

### C. A Frequency Domain Receiver Design Procedure

In this section, a frequency domain design of the equalizer is given. Again the Nyquist problem is considered using partial response technique, while the intersymbol interference due to the cubic term is taken into account in the design problem. The

actual filter frequency response is shown to represent a more practical solution and has superior noise performance compared to a standard PAM system.

The filter can be designed for an input signal which is a sum of the linear and cubic terms. This has one major drawback: the relative amplitude of the linear and cubic terms are not invariant, they are a function of the nonlinearity, signal amplitude and noise parameter. By using the method initially introduced by Gibby and Smith<sup>5</sup> in derivation of spectral requirement associated with the Nyquist problem, it was shown that

$$r_k = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} \sum_{n=-\infty}^{\infty} R(u + \frac{2n\pi}{T}) e^{jukT} du \quad (3)$$

$$\sum_n R_n(u) = T \sum_k r_k e^{-jukT} \quad (4)$$

where  $R(\omega)$  is the overall frequency response and  $r(t)$ , the corresponding time response and  $r_k = r(kT)$  for  $k$  an integer.

At the output of the nonlinearity, the linear part of the signal has a spectrum  $G_A(\omega)$  and the cubic part has a spectrum  $G_B(\omega)$ , the receiver filter frequency response is shown by  $G_R(\omega)$ , then for a partial response system one obtains

$$\sum_{n=-\infty}^{\infty} G_A^{(n)}(u) G_R^{(n)}(u) = T \cdot \sum_k r_k e^{-jukT} \quad (5)$$

$$\sum_{n=-\infty}^{\infty} G_B^{(n)}(u) G_R^{(n)}(u) = T \cdot \sum_k r_k e^{-jukT} \quad (6)$$

where

$$G_A^{(n)}(u) = G_A(u + \frac{2n\pi}{T}) \quad |u| \leq \pi/T$$

$$G_B^{(n)}(u) = G_B(u + \frac{2n\pi}{T}) \quad |u| \leq \pi/T$$

Subject to Eqs. (5) and (6) one desires to minimize output noise power at the receiver, i.e.,

$$\sigma_n^2 = \frac{N_0}{2\pi} \int_{-\pi/T}^{\pi/T} \sum_n |G_R^{(n)}(u)|^2 du \quad (7)$$

Thus, for each value of  $u$  in  $|u| \leq \pi/T$  one obtains the stationary point of

$$J = \sum_n \{ |G_R^{(n)}(u)|^2 + \lambda(u) G_A^{(n)}(u) G_R^{(n)}(u) + \nu(u) G_B^{(n)}(u) G_R^{(n)}(u) \}$$

from which one arrives at

$$G_R^{(n)}(u) = -\frac{\lambda(u)}{2} [G_A^{(n)}(u)]^* - \frac{\nu(u)}{2} [G_B^{(n)}(u)]^* \quad (8)$$

where  $[ ]^*$  denotes complex conjugate. Using Eq. (8) in Eqs. (5) and (6) yields

$$\sum_{n=-\infty}^{\infty} \left\{ -\frac{\lambda(u)}{2} |G_A^{(n)}(u)|^2 - \frac{\nu(u)}{2} G_A^{(n)}(u) [G_B^{(n)}(u)]^* \right\} = T \cdot \sum_k r_k e^{-jukT} \quad (9)$$

$$\sum_{n=-\infty}^{\infty} \left\{ -\frac{\lambda(u)}{2} [G_A^{(n)}(u)]^* G_B^{(n)}(u) - \frac{\nu(u)}{2} |G_B^{(n)}(u)|^2 \right\} = T \cdot \sum_k r_k e^{-jukT} \quad (10)$$

solving Eq. (10) for  $\nu(u)$

$$\nu(u) = \frac{-\left\{ \sum_n \lambda(u) [G_A^{(n)}(u)]^* G_B^{(n)}(u) \right\} - 2T \cdot \sum_k r_k e^{-jukT}}{\sum_n |G_B^{(n)}(u)|^2} \quad (11)$$

Substituting in Eq. (9) for  $\nu(u)$  and solving for  $\lambda(u)$  yields

$$\lambda(u) = \frac{2T \sum_k r_k e^{-jukT} - 2T \sum_k r_k e^{-jukT} \frac{\sum_n G_A^{(n)}(u) [G_B^{(n)}(u)]^*}{\sum_n |G_B^{(n)}(u)|^2}}{\sum_n \left\{ |G_A^{(n)}(u)|^2 + \left[ \frac{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)}{\sum_i |G_B^{(i)}(u)|^2} \right] G_A^{(i)}(u) [G_B^{(i)}(u)]^* \right\}} \quad (12)$$

and, similarly, one can find  $\nu(u)$ . Substituting for  $\lambda(u)$  and  $\nu(u)$  in Eq. (8) yields

$$G_R^{(n)}(u) = \frac{T \left\{ 1 - \frac{\sum_i G_A^{(i)}(u) [G_B^{(i)}(u)]^*}{\sum_i |G_B^{(i)}(u)|^2} \right\} [G_A^{(n)}(u)]^* \cdot \sum_k r_k e^{-jukT}}{\sum_n \left\{ |G_A^{(n)}(u)|^2 - \left[ \frac{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)}{\sum_i |G_B^{(i)}(u)|^2} \right] G_A^{(n)}(u) [G_B^{(n)}(u)]^* \right\}} +$$

$$+ \frac{T \left\{ 1 - \frac{\sum_i |G_A^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} \right\} [G_B^{(n)}(u)]^* \cdot \sum_k r_k e^{-jukT}}{\sum_n \left\{ G_A^{(n)}(u) \cdot [G_B^{(n)}(u)]^* - \left[ \frac{\sum_i |G_B^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} \right] G_A^{(n)}(u) |G_B^{(n)}(u)|^2 \right\}} \quad (13)$$

and

$$G_R(\omega) \triangleq \sum_{n=-\infty}^{\infty} G_R^{(n)}(\omega) \quad (14)$$

Equation (13) represents the design formula for a general pulse amplitude modulation system equalizer, including partial response signals. In the next section, specific design problems, involving the duobinary and Kretzmer's signals are discussed.

For simplicity, in subsequent use of Eq. (13) it is assumed that

$$G_R^{(n)}(u) \triangleq G_{R_s}^{(n)}(u) \cdot \sum_k r_k e^{-jukT} \quad (15)$$

where

$$G_{R_s}^{(n)}(u) = \frac{T \left\{ 1 - \frac{\sum_i G_A^{(i)}(u) [G_B^{(i)}(u)]^*}{\sum_i |G_B^{(i)}(u)|^2} \right\} [G_A^{(n)}(u)]^*}{\sum_n \left\{ |G_A^{(n)}(u)|^2 - \frac{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)}{\sum_i |G_B^{(i)}(u)|^2} G_A^{(n)}(u) [G_B^{(n)}(u)]^* \right\}} + \frac{T \left\{ 1 - \frac{\sum_i |G_A^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} \right\} [G_B^{(n)}(u)]^*}{\sum_n \left\{ G_A^{(n)}(u) \cdot [G_B^{(n)}(u)]^* - \frac{\sum_i |G_B^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} |G_A^{(n)}(u)|^2 \right\}} \quad (16)$$

#### D. Nonlinearity and Partial Response Systems

In this section, generalizations of some partial response signals which are of importance in practice are considered. The generalization involves the inclusion of nonlinearities in the system in order to suppress impulsive noise. Consider a duobinary system (an example of a simple partial response signal) for which

$$r_k = r(kT) = 0 \quad k \neq 0, 1$$

$$r_0 = r_1 = 1/2$$

where  $r_k$  is the sampled form of the overall time response. By using Eqs. (5) and (6) in a duobinary signaling scheme one obtains

$$\sum_{n=-\infty}^{\infty} G_A^{(n)}(u) G_R^{(n)}(u) = \frac{T}{2} (1 + e^{-juT}) \quad (17)$$

$$\sum_{n=-\infty}^{\infty} G_B^{(n)}(u) G_R^{(n)}(u) = \frac{T}{2} (1 + e^{-juT}) \quad (18)$$

and by following the same procedure as in the last section, results in

$$G_R^{(n)}(u) = \frac{\left\{ 1 - \frac{\sum_i G_A^{(i)}(u) [G_B^{(i)}(u)]^*}{\sum_i |G_B^{(i)}(u)|^2} \right\} \left[ \frac{T}{2} (1 + e^{-juT}) [G_A^{(n)}(u)]^* \right]}{\sum_n \left\{ |G_A^{(n)}(u)|^2 - \left[ \frac{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)}{\sum_i |G_B^{(i)}(u)|^2} \right] G_A^{(n)}(u) [G_B^{(n)}(u)]^* \right\}} + \frac{\frac{T}{2} \left\{ 1 - \frac{\sum_i |G_A^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} \right\} [G_B^{(n)}(u)]^* (1 + e^{-juT})}{\sum_n \left\{ G_A^{(n)}(u) \cdot [G_B^{(n)}(u)]^* - \left[ \frac{\sum_i |G_B^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} \right] \cdot |G_A^{(n)}(u)|^2 \right\}} \quad (19)$$

Equation (19) represents the receiver filter response for a duobinary system.

At this point, it is useful to discuss briefly the choice of the square root of a raised cosine signal as an input to the pulse amplitude modulation system.

The raised-cosine characteristic as commonly used in the literature on inter-symbol interference is given by

$$X(\omega) = \begin{cases} T & 0 \leq \omega \leq \frac{\pi}{T} (1 + \alpha) \\ \frac{T}{2} \left\{ 1 - \sin\left[\frac{T}{2\alpha} \left(\omega - \frac{\pi}{T}\right)\right] \right\} & \frac{\pi}{T} (1 - \alpha) \leq \omega \leq \frac{\pi}{T} (1 + \alpha) \end{cases}$$

where  $\alpha$  is the excess bandwidth parameter which is the amount of bandwidth used in excess of the Nyquist bandwidth divided by the latter. This  $\alpha$  is a variable less than one,  $\alpha = 0$  corresponds to the Nyquist bandwidth, and  $\alpha = 1$  corresponds to twice the Nyquist bandwidth.

The impulse response of the raised cosine is

$$x(t) = \frac{\sin(\frac{\pi T}{T})}{\frac{\pi t}{T}} \cdot \frac{\cos(\frac{\alpha \pi t}{T})}{(1 - 4\alpha^2 t^2 / T^2)}$$

which decreases asymptotically as  $1/t^3$  except for  $\alpha = 0$ .

$$\left[ x(t) = \frac{\sin(\pi t/T)}{(\pi t/T)} \text{ decreases as } 1/t \right]$$

The solution of Eq. (19) is shown in Figure 3. The time responses which result

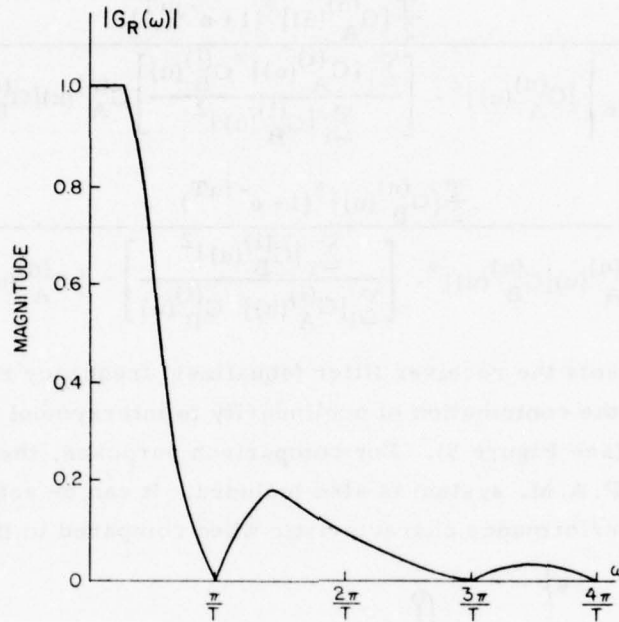


Fig. 3. The receiving filter frequency response for duobinary signaling with  $(1/2 + 1/2 D_L)$ .

from passing the linear and cubic terms through this filter are shown in Figure 4. If

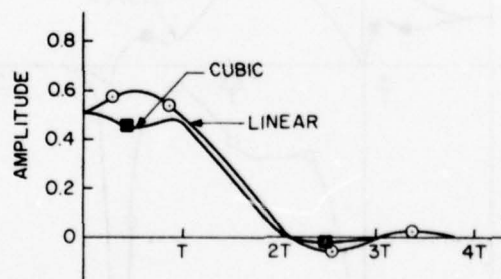


Fig. 4. Output time response of a duobinary system with  $(1/2 + 1/2 D_L)$ .

$$\sum_n G_B^{(n)}(u) \cdot G_R^{(n)}(u) = 0$$

which means that the contribution of nonlinearity to intersymbol interference is assumed to be zero. Considering this new constraint and following the same procedure as with Eqs. (5) through (13), one obtains

$$G_R(u) = \sum_n \left\{ \frac{\frac{T}{2} [G_A^{(n)}(u)]^* (1 + e^{-juT})}{\sum_n \left\{ |G_A^{(n)}(u)|^2 - \left[ \frac{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)}{\sum_i |G_B^{(i)}(u)|^2} \right] G_A^{(n)}(u) [G_B^{(n)}(u)]^* \right\}} + \frac{\frac{T}{2} [G_B^{(n)}(u)]^* (1 + e^{-juT})}{\sum_n \left\{ G_A^{(n)}(u) [G_B^{(n)}(u)]^* - \left[ \frac{\sum_i |G_B^{(i)}(u)|^2}{\sum_i [G_A^{(i)}(u)]^* G_B^{(i)}(u)} \right] \cdot |G_A^{(n)}(u)|^2 \right\}} \right\} \quad (20)$$

Equation (20) represents the receiver filter (equalizer) frequency response for a duobinary system when the contribution of nonlinearity to intersymbol interference is assumed to be zero (see Figure 5). For comparison purposes, the equalizer characteristic for a standard P.A.M. system is also included. It can be seen that the former has superior noise performance characteristic when compared to the latter.

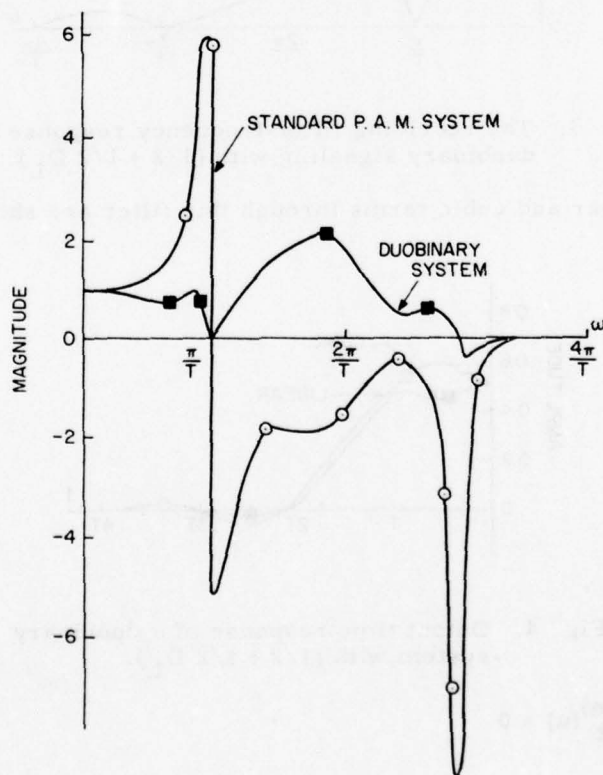


Fig. 5. Filter frequency response for the system with  $\sum_n G_B^{(n)}(u) \cdot G_R^{(n)}(u) = 0$ .

In this example  $\alpha = .5$  and

$$G_A(u) = \begin{cases} \sqrt{T} & 0 \leq u \leq \pi/2T \\ \sqrt{T/2} \cdot \{1 - \sin[T(u - \pi/T)]\} & \pi/2T \leq u \leq 3\pi/2T \end{cases}$$

$$G_B(u) = \begin{cases} \sqrt{T} & 0 \leq u \leq \frac{\pi}{2T} \\ T/2 \cdot \{1 - \sin[T(u - \pi/T)]\} \sqrt{T/2} \cdot \{1 - \sin[T(u - \pi/T)]\} & \frac{\pi}{2T} \leq u \leq \frac{3\pi}{2T} \end{cases}$$

Following the same procedure as for the duobinary system, the frequency response of the equalizer for the systems using other classes of partial response signals due to Kretzmer<sup>6</sup> can be obtained. For partial response signal with  $(1 + D_L)^2$ , where  $D_L = e^{-jut}$ , the corresponding receiver filter response becomes

$$G_R(u) = \sum_n G_{R_s}^{(n)}(u) \cdot [4 \cos^2[uT/2]] \quad (21)$$

where  $G_{R_s}(u)$  is given in Equation (16).

The equalizer frequency response of Eq. (21) is shown in Figure 6. The time

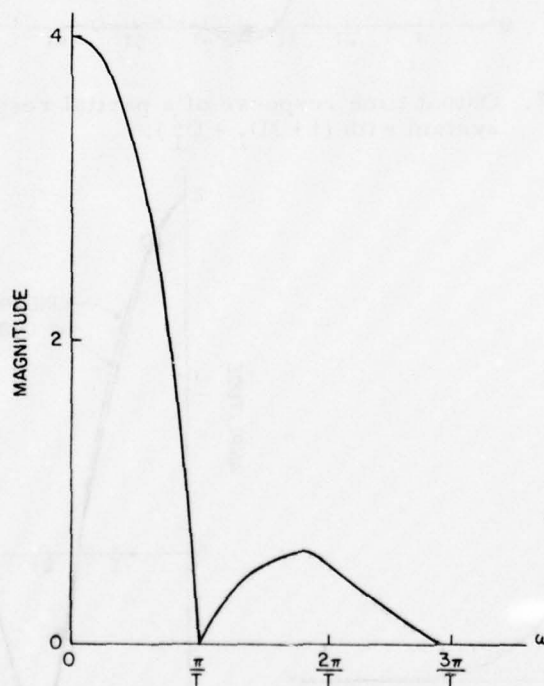


Fig. 6. The receiving filter frequency response for partial response system with  $(1 + D_L)^2$ .

response at the output of this filter due to linear and cubic terms are shown in Figure 7. For partial response signal with  $(2 + D_L - D_L^2)$ ,

$$G_R(u) = \sum_n G_{R_s}^{(n)}(u) \cdot \{ 2 + [\cos(uT) - \cos(2uT)] - j[\sin(uT) - \sin(2uT)] \} \quad (22)$$

which is shown in Figure 8. The output time responses of this filter are shown in Figure 9.

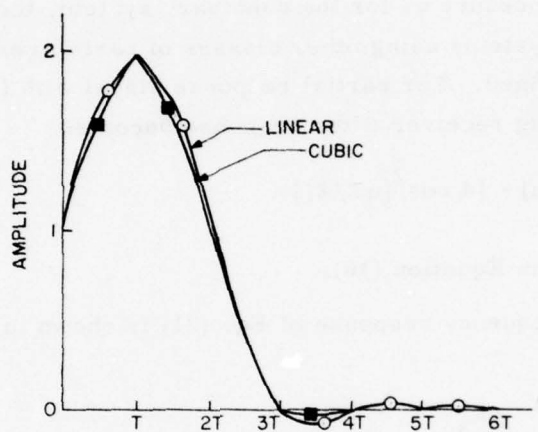


Fig. 7. Output time response of a partial response system with  $(1 + 2D_L + D_L^2)$ .

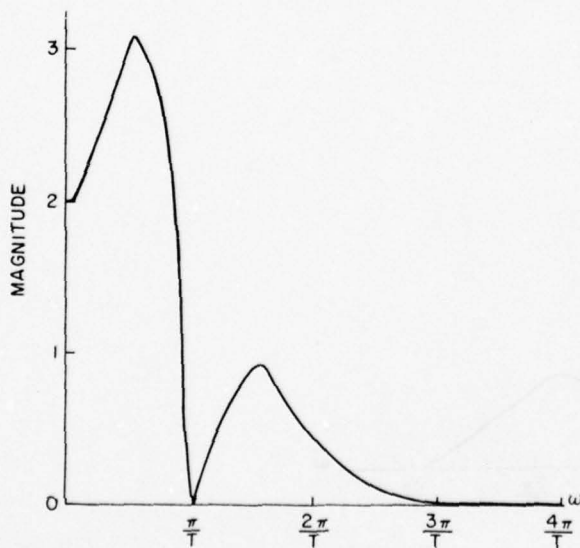


Fig. 8. The receiving filter frequency response for partial response system with  $(2 + D_L - D_L^2)$ .

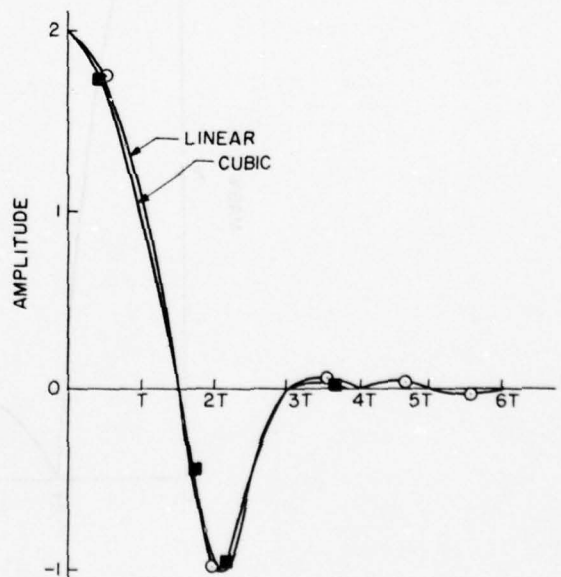


Fig. 9. Output time response of a partial response system with  $(2 + D_L - D_L^2)$ .

For partial response signal with  $(1 - D_L^2)$

$$G_R(u) = \sum_n G_R^{(n)}(u) \cdot [2 \sin(uT)] \quad (23)$$

which is shown in Figure 10. The output time responses of this filter are shown in Figure 11. For partial response signal with,  $(-1 + 2D_L^2 - D_L^4)$

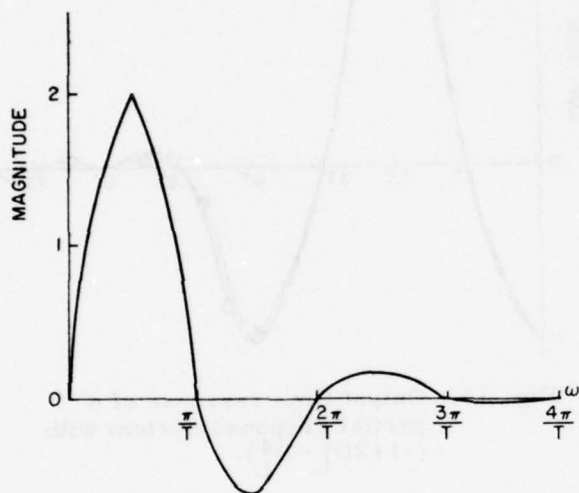


Fig. 10. The receiving filter frequency response for partial response signal with  $(1 - D_L^2)$ .

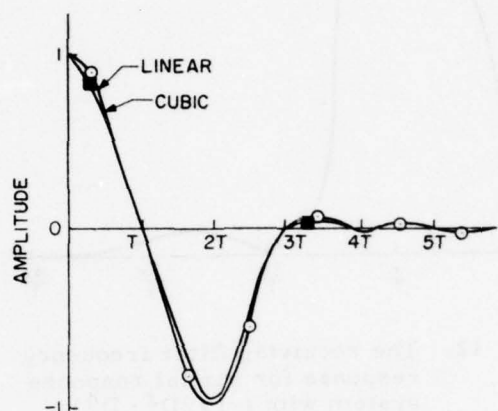


Fig. 11. Output time response of a partial response system with  $(1 - D_L^2)$ .

$$G_R(u) = \sum_n G_R^{(n)}(u) \cdot [4 \sin^2(uT)] \quad (23)$$

which is shown in Figure 12. The output time responses of this filter are shown in Figure 13. Comparing Figs. 6, 8, 10 and 12, it can be observed that the last model shown in Fig. 12 has only one spectral null within its pass-band, therefore from the viewpoint of implementation it is the best one. For comparison purposes, the output time response of the standard pulse amplitude modulation system with receiving filter frequency response shown in Fig. 5, is given in Figure 14. In this system, it was assumed that the intersymbol interference caused by the nonlinearity in system is zero. Comparing Fig. 14 with Figs. 7, 9, 11 and 13, it can be observed that, employing a controlled amount of intersymbol interference, results in well damped output time responses, which helps in increasing the timing jitter immunity.

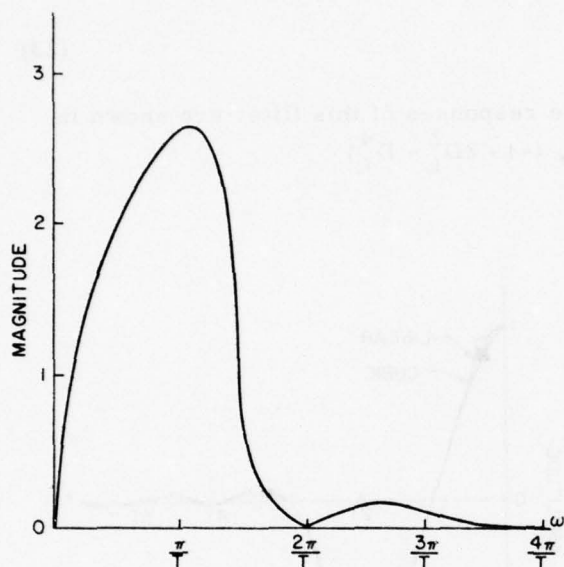


Fig. 12. The receiving filter frequency response for partial response system with  $(-1 + 2D_L^2 - D_L^4)$ .

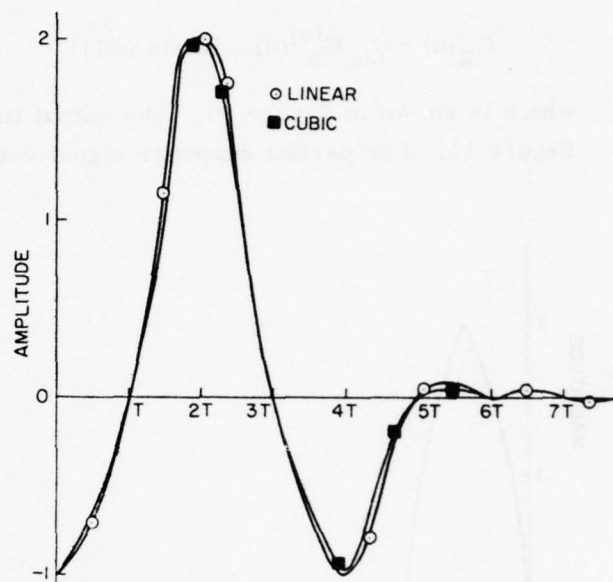


Fig. 13. Output time response of a partial response system with  $(-1 + 2D_L^2 - D_L^4)$ .

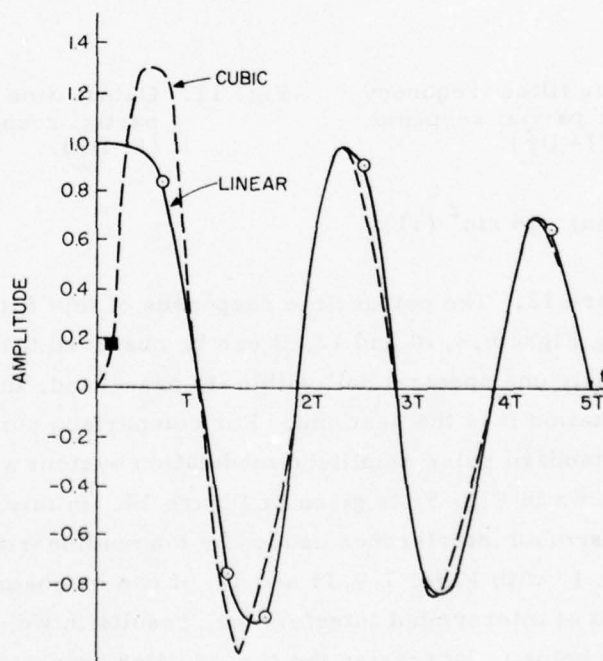


Fig. 14. Output time response of a standard P.A.M. system with no intersymbol interference.

This is another advantage of using partial response signals.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

M. Kavehrad and L. Kurz

#### REFERENCES

1. L. Kurz and G. Soloway, "The Nyquist Problem in the Presence of Nonlinearities in the Data Transmission System," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 322-327 (1976).
2. L. Kurz and M. Wernicki, "The Nyquist Problem in the Presence of Nonlinearities in the Duobinary System," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77, pp. 358-363 (1977).
3. S.S. Rappaport and L. Kurz, "An Optimal Nonlinear Detector for Digital Data Transmission Through Non-Gaussian Channels," IEEE Trans. Comm. Technology, COM-14, pp. 266-274 (June 1966).
4. R.W. Lucky, J. Salz and E.J. Weldon, Principles of Data Communication, (New York: McGraw-Hill, 1968).
5. R.A. Gibby and J.W. Smith, "Some Extensions of Nyquist's Telegraph Transmission Theory," BSTJ, 44, No. 7, pp. 1487-1510 (1965).
6. E.R. Kretzmer, "Generalization of a Technique for Binary Data Communication," IEEE Trans. Comm. Technology, COM-14, No. 1, pp. 67-68 (1966).

## ROBUSTIZED SEQUENTIAL PARTITION DETECTORS

H. S. Ashtiani and L. Kurz

In this report, an attempt will be made to robustize the performance of sequential partition detectors<sup>1</sup> against impulsive noise. In these detectors, although the test statistic upon which the decision is made is robust by itself in the sense of insensitivity to variations of the noise distribution in a severe impulsive noise environment, the contribution of the large amplitudes of the impulsive noise to the test statistic may reduce the effectiveness of its robustness. In these situations the performance of the detector may be improved by introducing regions of large signal ambiguity through setting two upper thresholds in the detector. The samples that fall into these regions are assumed to have come from the impulsive noise component of the assumed mixture noise, and are disregarded. The mixture noise is represented by  $n(t) = (1 - \epsilon)n_g(t) + \epsilon n_i(t)$  where  $n_g(t)$  and  $n_i(t)$  are the Gaussian and the impulsive noise components respectively. The region between the two thresholds in the sample space is partitioned into  $m$ -Intervals, the number of samples falling into each interval weighted appropriately are added, the sum is the test statistic upon which the detector bases its decision. In sequential partition detectors the decision, as to whether to accept or reject the presence of a signal in the received samples, is made sequentially using Wald's sequential procedure. In this procedure, samples are taken one at a time. The test stops when the received samples give adequate information for making a reliable decision. Wald's sequential procedure, called the sequential probability ratio test, is optimum in the sense of minimizing the average risk, when cost per observation is constant.

The sequential partition detectors are created as the result of applying Wald's sequential procedure to the  $m$ -Interval partition detectors.  $m$ -Interval partition detectors are robust in the sense that they are relatively insensitive to variations in the underlying noise distributions. Moreover, partition detectors can be made non-parametric (i.e., have fixed false alarm rate) by choosing the partitions under the hypothesis. The optimum property of the SPRT combined with properties of  $m$ -Interval partition detectors are the motivation behind considering the SPD for operation in an environment of mixed noise distribution.

It is assumed that the reader is familiar with the material of Ref. 1, including its terminology and notation. In Section A a robustized version of the sequential partition detector, and in Section B robustized versions of truncated sequential partition detectors are presented and their performance is analyzed.

### A. Robustized Sequential Partition Detector (RSPD)

We would like to desensitize the SPD to impulsive noise. For this purpose, the noise is represented by a mixture distribution model of the form

$$F(x) = (1 - \epsilon) P(x) + \epsilon H(x)$$

where  $P(x)$  is assumed to be a distribution having low variance representing the background Gaussian noise,  $H(x)$  is a large variance distribution representing large excursions of impulsive noise, and  $\epsilon$ , the mixing parameter, is assumed to be small, that is, we assume that the impulsive noise is infrequent.

An intuitively appealing way of combating the large amplitudes of impulsive noise is to place an upper limit on the acceptable sample value; in other words, we may limit the space of observation on both sides and disregard samples that fall beyond two upper and lower thresholds  $T_U$  and  $T_L$ . Rappaport and Kurz<sup>2</sup> explored this idea and showed that the rejection of samples that are too large compared to thresholds  $T_L$  and  $T_U$  improves the information rate of transmission in an impulsive noise environment. We will use the optimal thresholds established in Ref. 2 as a means of robustizing the performance of the SPD.

Since the impulsive noise component is infrequent, we will choose optimal partitions for Gaussian noise alone. As is shown in Ref. 3, quantiles of the noise distribution provide a near optimal partitioning of the sample space that can be easily implemented. Furthermore, robust quantile estimation methods are available<sup>4</sup> making this partitioning quite attractive in the impulsive noise environment under consideration.

#### RSPD Formulation

Let  $x_1, x_2, \dots, x_n$  be  $n$  i.i.d. r.v. from the c.d.f.  $F_v(x)$ ,  $v=0,1$ , let  $\vec{a} = (a_0 = T_L, a_1, \dots, a_{m-1}, a_m = T_U)$  be an ordered vector of constants subdividing the real axis from  $T_L$  to  $T_U$  into  $m$  intervals.

Define  $n_k$  as the number of samples falling into the interval

$$I[a_{k-1} < x_i \leq a_k, k = 1, \dots, m, i = 1, 2, \dots, n]$$

Let

$$P_{vk} = P_v[a_{k-1} < x_i \leq a_k] = F_v(a_k) - F_v(a_{k-1})$$

$v=0,1$ , represent the probability of a sample,  $x_i$ , falling into the  $k^{\text{th}}$  interval under the hypothesis or the alternative, under the constraint that

$$\sum_{k=1}^m P_k = C_U - C_L, \quad \text{and} \quad \sum_{k=1}^m n_k = n - n_\epsilon$$

where  $n_\epsilon$  is the number of samples that fall beyond the two thresholds  $T_L$  and  $T_U$  and are censored,  $C_U = F_0(T_U)$ , and  $C_L = F_0(T_L)$ .

Following the same steps as in Ref. 1, we form the test statistic

$$T_n = \sum_{k=1}^m b_k n_k, \quad b_k = \ln \frac{P_{ik}}{P_{0k}}$$

Decision procedure at the receiver is as for the case of SPD, i.e.,

"Continue sampling as long as,  $\ln B = b < T_n < a = \ln A$

Stop sampling, accepting hypothesis if,  $T_n \leq b$

Stop sampling, accepting alternative if,  $T_n \geq a$ "

where

$$A = \frac{1-\beta}{\alpha} \quad \text{and} \quad B = \frac{\beta}{1-\alpha}$$

as before.

Assuming the common case of shift of the mean alternative, we have

$$P_{0k} = F(a_k) - F(a_{k-1})$$

$$P_{ik} = F(a_k - \theta_1) - F(a_{k-1} - \theta_1) \approx P_{0k} + \theta_1 [f(a_{k-1}) - f(a_k)]$$

$$b_k = \ln \frac{P_{ik}}{P_{0k}} = \ln(1 + \theta_1 \frac{A_k}{P_{0k}}), \quad A_k = f(a_{k-1}) - f(a_k)$$

However, in order to establish  $T_n$  we have to estimate two additional quantiles  $F(a_0) = C_L$  and  $F(a_m) = C_U$ , as well as  $f(a_0)$  and  $f(a_m)$ .

The performance measures are the same as in Reference 1.

$$P_R(\theta) = \frac{1 - e^{-bt_0(\theta)}}{e^{-at_0(\theta)} - e^{-bt_0(\theta)}}$$

$$E_R(n/\theta) = \frac{1}{E(T_i)} \frac{b(e^{-at_0(\theta)} - 1) + a(1 - e^{-bt_0(\theta)})}{e^{-at_0(\theta)} - e^{-bt_0(\theta)}}, \quad E(T_i) \neq 0$$

$$E_R(n/\theta) = -\frac{b}{\sigma_{T_i}^2}, \quad E(T_i) = 0$$

$E(T_i)$  and  $\sigma_{T_i}^2$  are given in Reference 1.

$$E(T_i) = \sum_{k=1}^m b_k P_k = \theta_i \sum_{k=1}^m A_k + (\theta \theta_1 - \frac{1}{2} \theta_i^2) \sum_{k=1}^m \frac{A_k^2}{P_{0k}}$$

$$\sigma_{T_i}^2 = \theta_1^2 \sum_{k=1}^m A_k^2 / P_{0k}$$

where

$$\begin{aligned} \sum_{k=1}^m A_k &= \sum_{k=1}^m [f(a_{k-1}) - f(a_k)] \\ &= f(a_0) - f(a_1) \\ &+ \\ &\dots \\ &+ f(a_{m-1}) - f(a_m) \\ &= f(a_0) - f(a_m) \end{aligned}$$

$t_0(\theta)$  is +1 and -1 for  $\theta_0$  and  $\theta_1$ , respectively; and for small values of  $\theta$

$$t_0(\theta) = -\frac{2E(T_i)}{\sigma_{T_i}^2} = (1 - 2\frac{\theta}{\theta_1}) + \frac{2[f(a_m) - f(a_0)]}{\theta_1 \sum_{k=1}^m \frac{A_k^2}{P_{0k}}}$$

Thus,

$$\begin{aligned} E_R(n/\theta) &= \frac{1}{\theta_1 [f(a_0) - f(a_m)] + (\theta \theta_1 - \frac{1}{2} \theta_1^2) \sum_{k=1}^m \frac{A_k^2}{P_{0k}}} \\ &\cdot \frac{b(e^{at_0(\theta)} - 1) + a(1 - e^{bt_0(\theta)})}{e^{at_0(\theta)} - e^{bt_0(\theta)}}, \quad E(T_i) \neq 0 \end{aligned}$$

$$E_R(n/\theta) = - \frac{ab}{\theta_1^2 \sum_k^m \frac{A_k^2}{P_{0k}}} , \quad E(T_1) = 0$$

If  $\theta_0 = 0$  and  $\theta_1 = \theta_1$  then

$$E_R(n/\theta_0) \simeq \frac{|b|}{\theta_1 [f(a_m) - f(a_0)] + \frac{1}{2} \theta_1^2 \sum_k^m \frac{A_k^2}{P_{0k}}} \quad (1)$$

$$E_R(n/\theta_1) \simeq \frac{a}{\theta_1 [f(a_0) - f(a_m)] + \frac{1}{2} \theta_1^2 \sum_k^m \frac{A_k^2}{P_{0k}}} \quad (2)$$

#### B. Robustized Truncated Sequential Partition Detectors (RTSPD)

In order to have practical meaningful decision time, it is desirable to set a limit on the sample size, i.e., to truncate RSPD at the  $N_T^{\text{th}}$  stage of the sequential procedure. This would lead to a truncated version of RSPD. As in the case of TSPD, we choose gradually closing boundaries instead of abrupt truncation. The price of truncation is higher probabilities of error as in the case of TSPD.

##### RTSPD with Sloping Boundaries (RTSPD (SB))

Let RSPD be formulated as in Section except now let the stopping boundaries be given by  $a(1 - \frac{n}{N_T})$  and  $b(1 - \frac{n}{N_T})$ , with  $a$  and  $b$  as before. Following similar steps as for TSPD (SB) (see Ref. 1), we obtain the performance measures of RTSPD(SB) for shift of the mean alternative as follows

$$P_{RT(SB)}(\theta) = \frac{1 - e^{-|b|t_0(\theta)} E^*(e^{|b|n/N_T t_0(\theta)})}{e^{at_0(\theta)} E^{**}(e^{-an/N_T t_0(\theta)}) - e^{-|b|t_0(\theta)} E^*(e^{|b|n/N_T t_0(\theta)})}$$

$$E_{RT(SB)}(n/\theta) = \frac{1 - E_{RT}(n/\theta)/N_T}{\theta_1 [f(a_0) - f(a_m)] + (\theta\theta_1 - \frac{1}{2} \theta_1^2) \sum_k^m \frac{A_k^2}{P_{0k}}}$$

$$\cdot \frac{|b| [e^{at_0(\theta)} E^{**}(e^{-an/N_T t_0(\theta)}) - 1] + a [1 - e^{-|b|t_0(\theta)} E^*(e^{|b|n/N_T t_0(\theta)})]}{e^{at_0(\theta)} E^{**}(e^{-an/N_T t_0(\theta)}) - e^{-|b|t_0(\theta)} E^*(e^{|b|n/N_T t_0(\theta)})}$$

where  $E^*$  and  $E^{**}$  denote conditional expected values given  $H_0$  and  $H_1$ , respectively, and  $t_0(\theta)$ ,  $A_k$  and  $P_{0k}$  are as in Section A.

If  $\theta_0 = 0$  and  $\theta_1 = \theta_1$ , noting the fact that  $t_0(\theta)$  is still +1 and -1 for  $\theta_0$  and  $\theta_1$ , respectively, we obtain

$$E_{RT(SB)}^{(n/\theta_0)} \simeq \frac{|b|}{\theta_1 [f(a_m) - f(a_0)] + \frac{1}{2} \sum_k^m \frac{A_k^2}{P_{0k}} + \frac{|b|}{N_T}} \quad (3a)$$

$$E_{RT(SB)}^{(n/\theta_1)} \simeq \frac{a}{\theta_1 [f(a_0) - f(a_m)] + \frac{1}{2} \theta_1^2 \sum_k^m \frac{A_k^2}{P_{0k}} + \frac{a}{N_T}}$$

and

$$\alpha_{N_T(SB)} \simeq P_{RT(SB)}(\theta_0) \simeq e^{-a} (1 + a E_{RT(SB)}^{(n/\theta_0)}/N_T) \quad (3b)$$

$$\beta_{N_T(SB)} \simeq 1 - P_{RT(SB)}(\theta_1) \simeq e^{-|b|} (1 + |b| E_{RT(SB)}^{(n/\theta_1)}/N_T)$$

#### RTSPD with Curved Boundaries (RTSPD(CB))

Let RSPD be formulated as in Section A, except now let the boundaries be given by  $ae^{-rn/N_T}$  and  $be^{-rn/N_T}$  with,  $r$ , a constant representing rate of time-risk compensation and  $a$ ,  $b$ ,  $N_T$  defined as before.

Following similar steps as for RTSPD(SB) we find the ASN and errors of RTSPD(CB) for shift of the mean alternative. With  $\theta_0 = 0$  and  $\theta_1 = \theta_1$ , the expressions are

$$E_{RT(CB)}^{(n/\theta_0)} \simeq \frac{|b|}{\theta_1 [f(a_m) - f(a_0)] + \frac{1}{2} \theta_1^2 \sum_k^m \frac{A_k^2}{P_{0k}} + r \frac{|b|}{N_T}} \quad (4a)$$

$$E_{RT(CB)}^{(n/\theta_1)} \simeq \frac{a}{\theta_1 [f(a_0) - f(a_m)] + \frac{1}{2} \theta_1^2 \sum_k^m \frac{A_k^2}{P_{0k}} + r \frac{a}{N_T}}$$

$$\begin{aligned}\alpha_{N_T(CB)} &= P_{RT(CB)}(\theta_0) \simeq e^{-a} (1 + r a E_{RT(CB)}(n/\theta_0)/N_T) \\ \beta_{N_T(CB)} &= 1 - P_{RT(CB)}(\theta_1) \simeq e^{-|b|} (1 + r |b| E_{RT(CB)}(n/\theta_1)/N_T)\end{aligned}\quad (4b)$$

### C. Performance Curves and Concluding Remarks

The performance of the robustized sequential detectors of this report in terms of the average sample size as a function of signal-to-noise ratio, when the average probability of error is fixed, are given in Figure 1. The two types of errors are assumed equal, i. e.,  $\alpha = \beta = P_e/2$ . This leads to symmetrical stopping boundaries,  $a = \ln A = -b = -\ln B$ . Two error rates are considered,  $P_e = .01$  and  $P_e = .0001$ . The mixing parameter is assumed  $\epsilon = 10\%$ . For simplicity, the number of partitions are assumed  $m = 2$  and  $m = 4$ . For  $m = 2$ , the optimal partition is at  $a_1 = 0$ . This choice leads to a sign detector. For  $m = 4$ , near optimal partitions, obtained by equiprobable partitioning of the sample space, are selected. The impulsive noise thresholds are also assumed symmetrical, i. e.,  $a_m = T_U = -T_L = -a_0$ . Optimal thresholds of Ref. 2 are used; however, the choice of these thresholds has a negligible effect on the ASN. This is because  $[f(a_m) - f(a_0)]$  is small, for any reasonable choice of impulsive noise thresholds, compared to

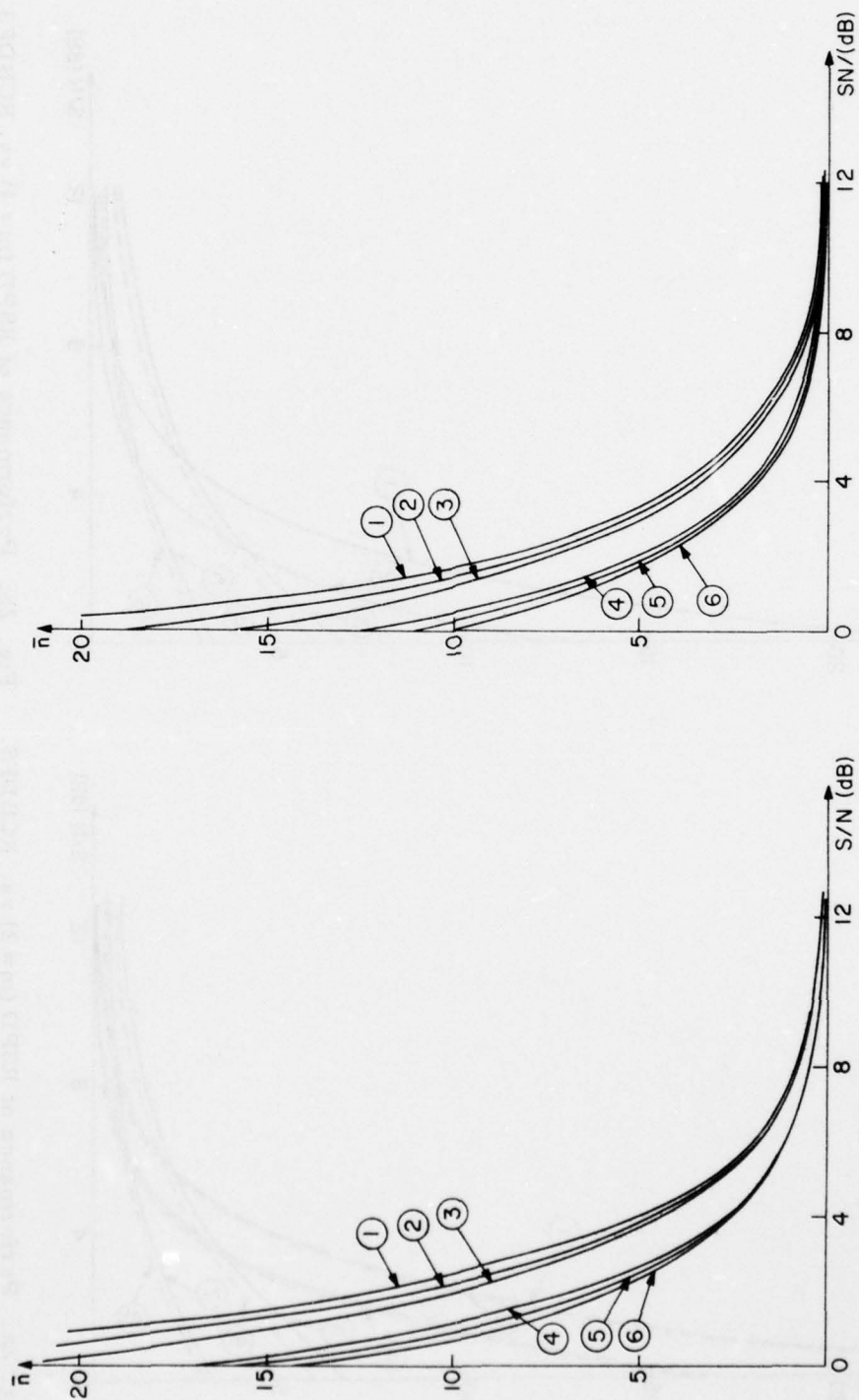
$$\sum_{k=1}^m \frac{A_k^2}{P_{0k}}$$

This means that protection against impulsive noise is achieved with no measurable increase in the ASN.

For the truncated RSPD, a truncation point of  $N_T = 100$  has been chosen. Since this choice is much larger than the ASN, the increase in the errors of the first and the second kind,  $\alpha, \beta$ , is negligible as can be seen from Equations (3b) and (4b). Thus, truncation has a negligible effect on the average probability of error.

For the purpose of comparison, the performance curves of non-truncated RSPD and truncated RSPD's with sloping and curved boundaries are given on the same graph. The rate of decay  $r = 2$  is assumed for the boundaries of RTSPD(CB). The average sample size for the RTSPD with curved boundaries is lower than the non-truncated RSPD and truncated RSPD with sloping boundaries.

A comparison between the non-truncated RSPD and its parametric competitor, RCBDIFS is shown in Figure 2. Two error rates,  $P_e = .01$  and  $P_e = .0001$  and two mixing parameters  $\epsilon = 5\%$  and  $\epsilon = 10\%$  are assumed with  $m = 2, 4$ . It is seen that the

Fig. 1a. Performance of RSPD ( $m = 2$ ).Fig. 1b. Performance of RSPD ( $m = 4$ ).

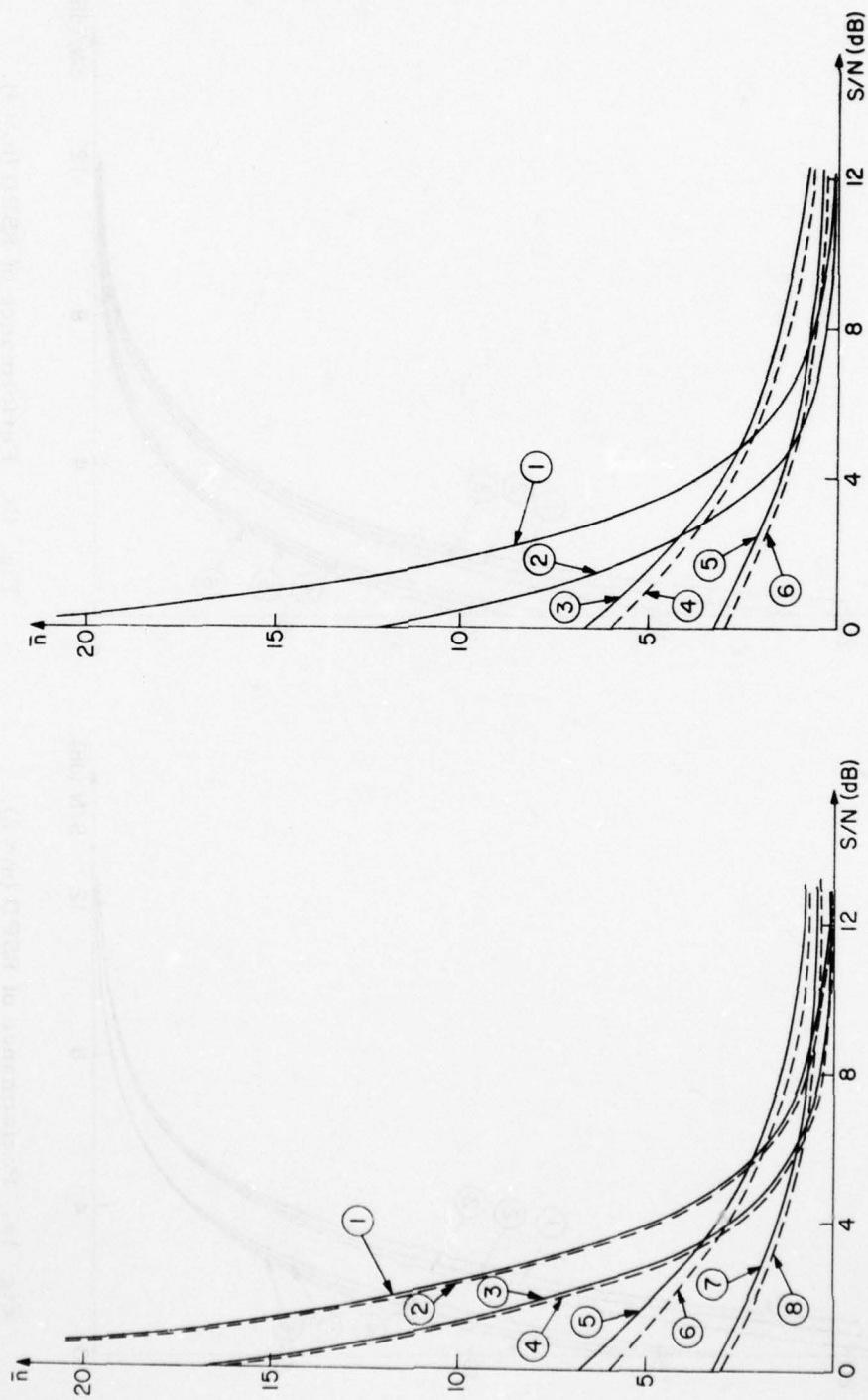


Fig. 2a. Performance of RSPD ( $m=2$ ) vs. RCBDFS. Fig. 2b. Performance of RSPD ( $m=4$ ) vs. RCBDFS.

RSPD is more robust than the RCBDFS. This was expected since RSPD is a nonparametric detector. Also, it is seen that RSPD is more efficient than the RCBDFS for higher S/N. The seemingly poor performance of RSPD at low signal-to-noise ratios is due to pronounced errors, especially since low values of  $m$  are considered. For large  $m$  this is eliminated because the SPD approaches the most powerful parametric detector.

The robustness of SPD is also verified by the fact that the performance of RCBDFS is significantly improved with respect to CBDFS while RSPD performs essentially the same as SPD.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

H. S. Ashitaini and L. Kurz

#### REFERENCES

1. Roger F. Dwyer and L. Kurz, "Sequential Partition Detectors," *Journal of Cybernetics*, Vol. 8, pp. 133-157 (1978).
2. S. S. Rappaport and L. Kurz, "Optimal Decision Thresholds for Digital Signaling in Non-Gaussian Noise," *IEEE Intl. Conf. Record*, Part 2, Vol. 13, pp. 198-212 (1965).
3. L. Kurz, "Nonparametric Detectors Based on Partition Tests," in *Nonparametric Methods in Communications*, P. Papantoni-Kazakos and D. Kazakos (eds.), pp 74-144, Marcel Dekker (1977).
4. P. Kersten and L. Kurz, "Robustized Robbins-Monro Algorithm with Applications to M-Interval Detection," *Information Sciences*, pp. 121-140 (October 1976).

## A REPRESENTATION THEORY AND ITS APPLICATION

R. Chassaing and L. Kurz

In this report, a new representation theory is developed to approximate the univariate and the bivariate p.d.f.'s in terms of orthogonal polynomials. The generalized Cauchy p.d.f. is chosen as reference to yield a set of orthogonal polynomials. Unlike the theory presented previously (see Ref. 1), the new approach does not require the Markov assumption for the underlying distributions. Using the results of the new representation, the autocorrelation function at the output of the nonlinearity specified in Fig. 1 of Ref. 1 is obtained. The representation methods of this report are useful in obtaining approximate expressions for likelihood ratios, in performing spectral analysis of nongaussian processes shaped by memoryless nonlinearities and related problems. The procedure outlined here may be easily extended to  $n$  dimensions.

A. One-Dimensional Representation Theory

A basic method of representing a probability density function (p.d.f.) is by a generalized Fourier Series,

$$p_D = \sum A_i \phi_i(x) p_R(x) \quad (1)$$

where  $p_D$  is the desired (unknown) p.d.f.,  $p_R$  a reference p.d.f. and  $\phi_i(x)$  are the orthogonal polynomials with respect to the weight function, or

$$\int_{-\infty}^{\infty} \phi_i(x) \phi_j(x) p_R(x) dx = \delta_{ij} \quad (2)$$

where  $\delta_{ij}$  is the Kronecker delta. The coefficients  $A_i$  can be found using Eq. (1) and the orthogonality property Eq. (2),

$$A_i = \int_{-\infty}^{\infty} \phi_i(x) p_D(x) dx \quad (3)$$

Since  $\phi_i(x)$  are polynomials,  $A_i$  is expressed in terms of the moments of  $p_D$ . Although Eq. (1) may or may not converge, such expansion is still useful in that only a few terms lead to a satisfactory approximation of an unknown p.d.f.,

$$p_D \approx \sum_{i=0}^N A_i \phi_i(x) p_R(x) \quad N \text{ small} \quad (4)$$

Using the approximation of Eq. (4), only the first few moments of  $p_D$  are required (these moments can be estimated).

The reference p.d.f. is chosen as the "generalized Cauchy" p.d.f.<sup>2,3</sup>

$$p_R \approx \frac{c\eta \nu^{-1/c} \Gamma(\nu+1/c)}{2\Gamma(1/c)\Gamma(\nu)(1 + \frac{[\eta |x|]^c}{\nu})^{\nu+1/c}} \quad (5)$$

where

$$\eta = \sigma^{-1} \left[ \frac{\Gamma(3/c)}{\Gamma(1/c)} \right]^{1/2} \quad (6)$$

since such reference p.d.f. Eq. (5) can represent a wide range of density functions. For  $c = 2$ ,  $\nu = 1/2$ ,  $p_R$  reduces to the Cauchy density, a p.d.f. having infinite variance. The p.d.f. of Eq. (5) can be used successfully to characterize impulsive noise<sup>4,5</sup> with the parameters  $c$  and  $\nu$  indicating the degree of impulsiveness. Such p.d.f.,  $p_R$ , can also characterize background Gaussian noise, since as  $\nu \rightarrow \infty$ ,  $c = 2$ ,  $p_R$  reduces to the Gaussian p.d.f.

As a special case, the desired p.d.f.  $p_D$ , is chosen to be the "generalized" Gaussian p.d.f.<sup>2</sup>

$$p_D = \frac{c\eta}{2\Gamma(1/c)} \exp \{-(\eta |x|)^c\} \quad (7)$$

For  $c = 2$ , Eq. (7) reduces to the Gaussian p.d.f., and for  $c = 1$ ,  $p_D$  becomes the Laplace p.d.f. The density of Eq. (5) can also represent impulsive noise by choosing  $.1 < c < .6$  (see Ref. 6), and therefore could be used as a reference p.d.f. for impulsive noise. However, the chosen reference p.d.f. of the class Eq. (5) represents a wider class of distribution in modeling impulsive noise.

#### B. Evaluation of the Representation Expansion

The polynomials  $\phi_n(x)$  can be found using a method developed by Szego,<sup>7</sup> namely

$$\phi_n(x) = (D_n D_{n-1})^{1/2} \cdot D x_n \quad (8)$$

where

$$D_n = \begin{vmatrix} C_0 & C_1 & \cdots & C_n \\ C_1 & \cdots & \cdots & C_{n+1} \\ \vdots & & & \\ C_n & C_{n+1} & \cdots & C_{2n} \end{vmatrix} \quad (9)$$

$$Dx_n = \begin{vmatrix} C_0 & C_1 & \cdots & C_n \\ C_1 & \cdot & \cdots & C_{n+1} \\ \vdots & & & \\ C_{n-1} & C_n & \cdots & C_{2n-1} \\ x^0 & x^1 & \cdots & x^n \end{vmatrix}$$

and the  $C_n$  are moments of the reference p.d.f. The moments of the desired and the reference p.d.f. are: for  $p_D$

$$G_{2r} = \frac{\Gamma(\frac{2r+1}{c})}{\eta^{2r} \Gamma(\frac{1}{c})}$$

and for  $p_R$

$$S_{2r} = \frac{h^r \Gamma(m - \frac{1}{2} - r) \Gamma(r + \frac{1}{2})}{\Gamma(m - \frac{1}{2}) \Gamma(\frac{1}{2})}$$

A computer program was developed to find the approximated p.d.f. resulting from the expansion Eq. (4) for several cases of reference and desired p.d.f.'s. The results are found in Tables I to IV.

### C. Two-Dimensional Representation Theory

The representation theory used in the previous two sections can be extended to approximate a two-dimensional p.d.f. Such p.d.f. can later be used to find the autocorrelation function at the output of a nonlinearity. The basic expansion has the form

$$p_D(x_1, x_2) = p_R(x_1) p_R(x_2) \sum_{m=0}^N \sum_{n=0}^N A_{mn} \phi_m(x_1) \phi_n(x_2) \quad (10)$$

where  $p_D(x_1, x_2)$  is the desired two-dimensional p.d.f.,  $\phi_m(x_1)$  and  $\phi_n(x_2)$  are the two sets of polynomials orthogonal to  $p_R(x_1)$  and  $p_R(x_2)$ , or

$$\begin{aligned} \int_{-\infty}^{\infty} \phi_i(x_1) \phi_j(x_1) p_R(x_1) dx_1 &= \delta_{ij} \\ \int_{-\infty}^{\infty} \phi_u(x_2) \phi_v(x_2) p_R(x_2) dx_2 &= \delta_{uv} \end{aligned} \quad (11)$$

$p_R(x_1)$  and  $p_R(x_2)$  are the one-dimensional p.d.f.'s given by Equation (5). The coefficients

TABLE II

x	Desired p.d.f.	Approximated p.d.f. (c = 4 for Reference p.d.f.)			
		$\nu = 6$	$\nu = 10$	$\nu = 20$	$\nu = 26$
0	Gaussian	.5743	.5849	.5909	.5921
.5		.5364	.5338	.5278	.5259
1.0		.4094	.3841	.3580	.3512
1.5		.2252	.2088	.1945	.1910
2.0		.0889	.0956	.1032	.1053
2.5		.0253	.02654	.0257	.0253
3.0		.0048	.00313	.00159	.00127
3.5		.00069	.000186	.0000257	.000013
4.0		.000089	.0000077	.0000002	.000000036
4.5		.000012	.0000003	.000000001	.00000000001

TABLE I

x	Desired p.d.f.	Approximated p.d.f. (c = 2 for Reference p.d.f.)			
		$\nu = 6$	$\nu = 10$	$\nu = 20$	$\nu = 26$
0	Gaussian	.5665	.6078	.6285	.6323
0.5		.4911	.5316	.5529	.5569
1.0		.3249	.3586	.3777	.3812
1.5		.1708	.1909	.2022	.2042
2.0		.0751	.0828	.0861	.0865
2.5		.0290	.0302	.02956	.0293
3.0		.0102	.0095	.00828	.00799
3.5		.0033	.0026	.00189	.00175
4.0		.0010	.000587	.000333	.0003
4.5		.00027	.000085	.0000347	.000033

Reference p.d.f.:  $\frac{K}{(x^2 + 2\nu)^{\nu+1/2}}$

AD-A063 181

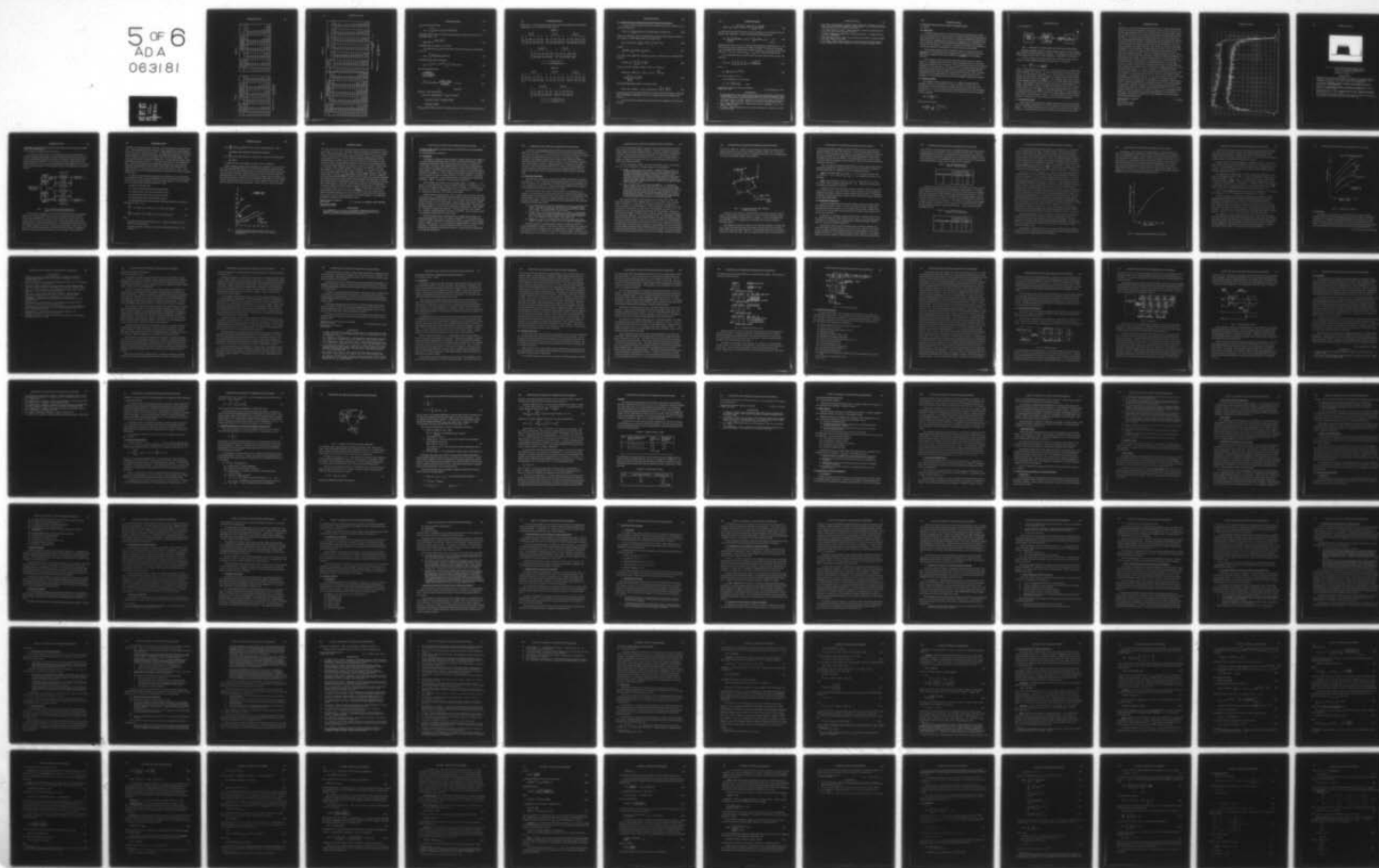
POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 9/3  
PROGRESS REPORT NUMBER 43 TO THE JOINT SERVICES TECHNICAL ADVIS--ETC(U)  
NOV 78 A A OLINER F44620-78-C-0074

UNCLASSIFIED

POLY-MRI-452.43-78

NL

5 OF 6  
ADA  
063181



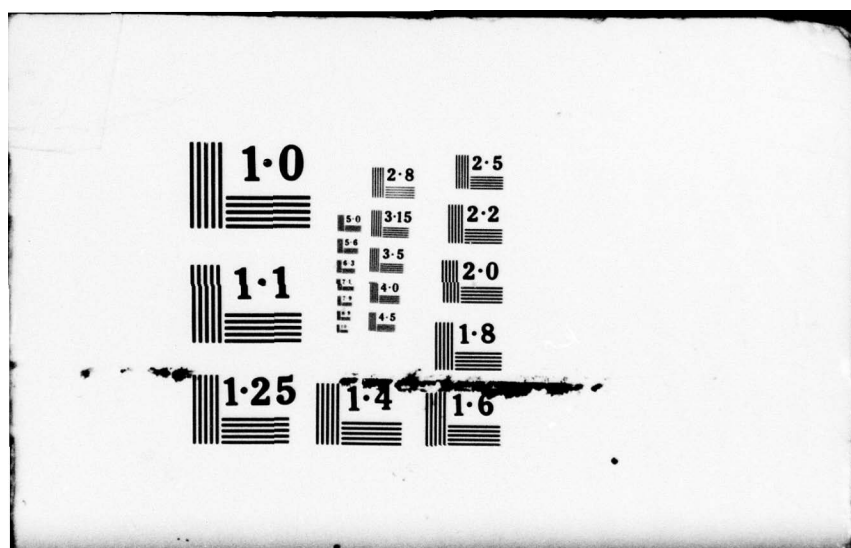


TABLE II

x	Desired p.d.f.	Approximated p.d.f. (c = 4 for Reference p.d.f.)			
		v = 6	v = 10	v = 20	v = 26
0	Gaussian	.5743	.5849	.5909	.5921
.5		.5364	.5338	.5278	.5259
1.0		.4094	.3841	.3580	.3512
1.5		.2252	.2088	.1945	.1910
2.0		.0889	.0956	.1032	.1053
2.5		.0253	.02654	.0257	.0253
3.0		.0048	.00313	.00159	.00127
3.5		.00069	.000186	.0000257	.000013
4.0		.000089	.0000077	.0000002	.00000036
4.5		.000016	.0000003	.000000001	.00000000001

TABLE I

x	Desired p.d.f.	Approximated p.d.f. (c = 2 for Reference p.d.f.)			
		v = 6	v = 10	v = 20	v = 26
0	Gaussian	.5665	.6078	.6285	.6323
0.5		.4911	.5316	.5529	.5569
1.0		.3249	.3586	.3777	.3812
1.5		.1708	.1909	.2022	.2042
2.0		.0751	.0828	.0861	.0865
2.5		.0290	.0302	.02956	.0293
3.0		.0102	.0095	.00828	.00799
3.5		.0033	.0026	.00189	.00175
4.0		.0010	.000587	.000333	.0003
4.5		.00027	.000085	.0000347	.000033

$$\text{Reference p.d.f.: } \frac{K}{(x^2 + 2v)^v + 1/2}$$

TABLE IV

x	Desired p.d.f.	Approximated p.d.f. (c = 2 for Reference p.d.f.)					Approximated p.d.f.				
		v = 10	v = 12	v = 14	v = 18	v = 20	v = 6	v = 10	v = 14	v = 18	v = 26
0	.7071	.6079	.6155	.6204	.6285	.6323	.5665	.6079	.6205	.6265	.6323
.5	.3486	.5127	.5155	.5165	.5163	.5163	.4494	.3921	.3371	.2975	.2468
1.0	.1719	.3117	.3065	.3012	.2884	.2884	.2189	.0119	-.1397	-.2431	-.3710
1.5	.0848	.1430	.1358	.1296	.1161	.1161	.0544	-.1639	-.3071	-.4001	-.5113
2.0	.0418	.0550	.0514	.0487	.0437	.0437	-.00487	-.1232	-.1832	-.2162	-.2502
2.5	.0206	.0216	.0216	.0219	.0233	.0233	-.00917	-.0336	-.0296	-.0212	-.0062
3.0	.01016	.0103	.0115	.0125	.0149	.0149	-.00129	.0155	.0364	.0517	.0706
3.5	.0050	.0057	.0066	.0072	.0084	.0084	.00387	.0254	.0396	.0474	.0548
4.0	.00247	.0031	.0035	.0037	.0039	.0039	.00520	.0195	.0248	.0263	.0263
4.5	.0012	.00166	.0017	.00169	.0015	.0015	.00465	.0118	.0122	.0114	.0097
5.0	.0006	.000836	.000781	.000706	.00051	.00051	.00354	.0063	.0053	.0043	.00298
5.5	.000296	.0004	.000338	.000277	.000162	.000162	.00249	.0031	.0021	.00145	.00081
6.0	.000146	.000192	.000142	.00010	.0000475	.0000475	.00169	.0015	.00081	.00046	.00020
6.5	.000072	.0000905	.000058	.0000385	.0000133	.0000133	.001127	.00071	.00030	.00014	.000046
7.0	.0000355	.0000425	.0000241	.000014	.0000036	.0000036	.000746	.000337	.00011	.000042	.000010

Desired p.d.f. =  $K \exp \left\{ -\frac{2}{\sigma} \frac{6}{1/2} \left| \frac{x}{1/2} \right|^{1/2} \right\}$ ,  $c = 1/2$

Reference p.d.f. =  $\frac{K}{(x^2 + 2v)^{v+1/2}}$ ,  $c = 2$

$A_{mn}$  can be evaluated using

$$A_{mn} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_D(x_1, x_2) \phi_m(x_1) \phi_n(x_2) dx_1 dx_2 \quad (12)$$

Using  $p_R(x)$  given by Eq. (5) yields the same orthogonal polynomials as for the one-dimensional case. Let

$$p_D(x_1, x_2) = a e^{-b(x_1^2 + x_2^2)^{1/s}} \quad (13)$$

Normalizing Eq. (13) yields  $a = b^s / s\pi \Gamma(s)$ .

The coefficients  $A_{mn}$  can be expressed in terms of the moments of the desired p.d.f., or,

$$A_{mn} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1^m x_2^n p_D(x_1, x_2) dx_1 dx_2 \quad (14)$$

Transforming into polar coordinates,

$$A_{mn} = a \int_0^{\infty} r^{m+n+1} e^{-br^{2/s}} dr \int_0^{2\pi} \cos^m \theta \sin^n \theta d\theta \quad (15)$$

The first integral becomes

$$\frac{\Gamma\left(\frac{m+n+2}{2}\right)^s}{\Gamma(s) 2\pi b\left(\frac{m+n}{2}\right)^s} \quad (16)$$

$$\int_0^{2\pi} \cos^m \theta \sin^n \theta d\theta = \frac{\Gamma\left(\frac{m+1}{2}\right) \Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{m+n+2}{2}\right)} \quad m, n \text{ even} \quad (17)$$

= 0

Otherwise

Letting  $N=4$ , Eq. (10) becomes

$$\begin{aligned} p_D(x_1, x_2) = & p_R(x_1) p_R(x_2) \{ 1 + a_{22} \phi_2(x_1) \phi_2(x_2) \\ & + a_{24} \phi_2(x_1) \phi_4(x_2) + a_{42} \phi_4(x_1) \phi_2(x_2) \\ & + a_{44} \phi_4(x_1) \phi_4(x_2) \} \end{aligned} \quad (18)$$

with  $A_{mn}$  given by Eqs. (16) and (17),  $\phi_1(x)$  given by Eqs. (8) and (9) and  $p_R(x)$  given by

Equation (5). A computer program was used to find the approximations of several two-dimensional p.d.f.'s, as shown in Tables V and VI.

TABLE V

$x_1 = 0$						$x_1 = .5$					$x_1 = 1.$				
$x_2$	0	.5	1.	1.5	2	0	.5	1.	1.5	2.	0	.5	1.	1.5	2.
$p_D$	.318	.248	.117	.034	.006	.248	.193	.091	.026	.004	.117	.091	.043	.012	.002
$p_A$	.159	.139	.093	.049	.022	.139	.122	.084	.045	.019	.093	.084	.061	.033	.012

$x_1 = 1.5$						$x_1 = 2.$				
$x_2$	0	.5	1.	1.5	2.	0	.5	1.	1.5	2
$p_D$	.033	.026	.012	.003	.0006	.005	.004	.002	.0006	.0001
$p_A$	.049	.045	.033	.018	.006	.02	.019	.012	.006	.003

$b = s = 1$  in Desired p.d.f.

$c = 2, v = 26$  in Reference p.d.f.

TABLE VI

$x_1 = 0$						$x_1 = .5$					$x_1 = 1.$				
$x_2$	0	.5	1.	1.5	2.	0	.5	1.	1.5	2.	0	.5	1.	1.5	2.
$p_D$	.159	.14	.096	.051	.02	.14	.12	.08	.04	.019	.09	.08	.058	.03	.013
$p_A$	.2	.15	.072	.017	.002	.15	.12	.07	.03	.01	.072	.07	.06	.04	.021

$x_1 = 1.5$						$x_1 = 2.$				
$x_2$	0	.5	1.	1.5	2.	0	.5	1.	1.5	2
$p_D$	.05	.045	.03	.016	.007	.021	.019	.013	.007	.003
$p_A$	.017	.028	.04	.038	.02	.002	.01	.02	.021	.01

$b = 1/2, s = 1$  in Desired p.d.f.

$c = 2, v = 6$  in Reference p.d.f.

#### D. Finding the Autocorrelation Function at the Output of a Nonlinearity

Using the transform method,<sup>1</sup> the autocorrelation function at the output of a memoryless nonlinearity is

$$R_y(t_1, t_2) = \frac{1}{(2\pi j)^2} \int_c f(w_1) dw_1 \int_c f(w_2) dw_2 M_s(w_1, w_2) M_n(w_1, w_2) \quad (19)$$

where  $M_s(\cdot)$  and  $M_n(\cdot)$  are the moment generating functions (the Laplace transform of the joint p.d.f.) of the signal and the noise respectively.

$$f(w) = f_-(w) + f_+(w) = \int_{-\infty}^0 g(x) e^{-wx} dx + \int_0^{\infty} g(x) e^{-wx} dx \quad (20)$$

From<sup>8</sup>

$$\int_C f(w) dw = \int_{C_-} f_-(w) dw + \int_{C_+} f_+(w) dw \quad (21)$$

The Taylor series expansion of  $\log M(w_1, w_2)$  has as coefficients the cumulants of the process,

$$\log M(w_1, w_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} w_1^i w_2^j \frac{\lambda_{ij}}{i! j!} \quad (22)$$

where  $\lambda_{ij}$  are the cumulants. Equation (22) can be written as,

$$M_n(w_1, w_2) = \exp[w_1 \lambda_{10} + w_2 \lambda_{01} + w_1 w_2 \lambda_{11} + \frac{w_1^2}{2} \lambda_{02} + \frac{w_2^2}{2} \lambda_{20} + \dots] \quad (23)$$

The moment generating function of Gaussian noise is

$$M_G(w_1, w_2) = \exp[w_1 n_1 + w_2 n_2 + w_1 w_2 R_n(t_1, t_2) + \frac{w_1^2}{2} \sigma_1^2 + \frac{w_2^2}{2} \sigma_2^2] \quad (24)$$

where  $R_n(t_1, t_2)$  is the correlation function,  $n_i$  the mean and  $\sigma_i^2$  the variance. Note that the first exponential expression in Eq. (23) corresponds then to the Gaussian noise process.

For the case of zero means and unit variances for the Gaussian noise, Eq. (23) becomes

$$M_n(w_1, w_2) = \exp\left[\frac{w_1^2 + w_2^2 + 2w_1w_2}{2}\right] \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{\phi_{ij} w_1^i w_2^j}{i!j!} \quad (25)$$

$\phi_{ij}$  were derived by Bowen,<sup>9</sup> however, there is a slight error in his expression (his indexing in the summation). It can be shown that the coefficients  $\phi_{ij}$  are

$$\phi_{ij} = \sum_{\alpha=1}^i \sum_{\beta=1}^j \binom{i-1}{\alpha-1} \binom{j}{\beta} \phi_{i-\alpha, j-\beta} \lambda_{\alpha\beta} + \begin{cases} \sum_{\alpha=3}^i \binom{i-1}{\alpha-1} \phi_{i-\alpha, j} \lambda_{\alpha 0} & i \geq 3 \\ 0 & 1 \leq i \leq 2 \end{cases} \quad (26)$$

Equation (26) is valid for  $i \geq 1, j \geq 0$ , with the double summation omitted when  $j = 0$ . The univariate form of Eq. (26) appears in Kendall and Stuart.<sup>3</sup> The bivariate cumulants  $\lambda_{mn}$  can be expressed in terms of the bivariate moments  $C_{mn}$  of the desired two-dimensional p.d.f. Eq. (13), (in this case  $C_{mn} = A_{mn}$  of Eq. (14) or using Equation (18).

For a nonlinearity with a transform  $f(w)$ , the autocorrelation function of Eq. (19) reduces to

$$R_y(t_1, t_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} h_L h_R \frac{\phi_{ij} A^{m+n} \rho^k}{i!j!k!m!n!} \quad (27)$$

where

$$h_L = \frac{1}{2\pi j} \int_C f(w) w^L e^{w^2/2} dw \quad (28)$$

for  $L = k+i+m$  and  $R = k+j+n$ .

If the nonlinearity is a hard limiter

$$h_L = (j)^{L-1} \frac{2^{L/2} a}{\pi} \Gamma\left(\frac{L}{2}\right) \quad L \text{ odd}$$

Joint Services Technical Advisory Committee  
F44620-74-C-0056

R. Chassaing and L. Kurz

#### REFERENCES

1. R. Chassaing and L. Kurz, "Autocorrelation Function at the Output of Memoryless Nonlinearities Using a Differential Equation Approach," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77, pp. 261-269 (1977).
2. J.M. Miller and J.B. Thomas, "Detectors for Discrete Time Signals in Non-Gaussian Noise," IEEE Trans. Info. Theory, Vol. IT-18, pp. 241-250 (March 1972).
3. M.G. Kendall and A.M. Stuart, "The Advanced Theory of Statistics," Vol. I, 2nd Edition, Hafner (1963).
4. P. Mertz, "Model of Impulsive Noise for Data Transmission," IRE Trans. Comm. Tech. (June 1961).

5. M.M. Hall, "A New Model for Impulsive Noise Phenomena: Application to Atmospheric Noise Communication Channels," Stanford Electr. Lab., Stanford University, Technical Report 3412-8 and 7050-7 (April 1966).
6. N.R. Algazi and R.M. Lerner, "Binary Detection in White Non-Gaussian Noise," Lincoln Lab. MIT, Technical Report D5-2138.
7. G. Szego, "Orthogonal Polynomials," American Math. Soc., Colloquium Publications, Vol. XXIII (1939).
8. W.B. Davenport and W.L. Root, "An Introduction to the Theory of Random Signals and Noise," McGraw Hill (1958).
9. B.A. Bowen, "The Transform Method for Nonlinear Devices with Non-Gaussian Noise," IEEE Trans. on Info. Theory, Vol. IT-13, pp. 326-328 (April 1967).

## A UNIFORM POWER SPECTRAL DENSITY JAMMING SIGNAL

F. Cassara

A. Introduction

It is often useful in electronic countermeasures to transmit high power noise over some prescribed band of frequencies in an attempt to interfere with transmissions from an unfriendly source. Since we do not know with any certitude the frequencies at which the source transmits and/or receives, it is desirable to use a high power signal with a continuous uniform power spectral density bandlimited over some frequency band for the noise jammer. In this work, a technique is described for generating such a signal with flexibility in designing for its center frequency and bandwidth.

The technique employed utilized Woodward's Theorem<sup>1</sup> which states that the spectrum of an FM signal with large modulation index ( $\beta_{rms} > 10$ ) takes on the same shape as the probability density function (pdf) of the amplitude of the modulating waveform.

A block diagram of the system used to generate the noise jammer is shown in Figure 1. A nonlinear network is used to transform Gaussian noise into a stochastic signal whose amplitude has a uniform pdf. The resultant signal is then used to frequency modulate a carrier with large  $\beta$ . The spectrum of the transmitted signal will then be uniform and continuous centered around the carrier frequency with bandwidth (BW) approximately equal to twice the peak frequency deviation (Carson's Rule). The power contained in the transmitted signal can be made large by using a high power FM modulator or by using efficient nonlinear RF power amplifiers following the modulator.

B. Nonlinear Network

The nonlinear network required in Fig. 1 is readily determined using the techniques relating to transformation of a random variable. If  $x$  (see Fig. 1) is assumed to be a zero mean stationary Gaussian random process with variance  $\sigma_x^2$ , and we desire  $y$  to be a zero mean uniformly distributed process over the normalized interval  $[-\frac{1}{2}, \frac{1}{2}]$ , then the transfer function of the nonlinear network can be shown<sup>2</sup> to take on the form

$$g(x) = \text{Erf}\left(\frac{x}{\sigma_x}\right) - \frac{1}{2} \quad (1)$$

where the error function

$$\text{Erf}\left(\frac{x}{\sigma_x}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x/\sigma_x} e^{-z^2/2} dz \quad (2)$$

is well tabulated.<sup>3</sup>

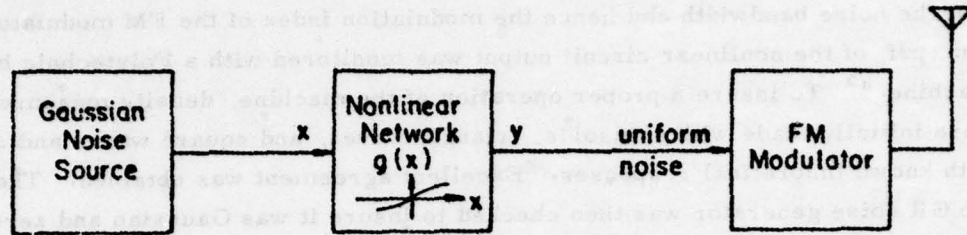


Fig. 1. Generation of a uniform power spectral density jamming signal.

Although a broadband nonlinear diode waveshaping network could be designed with transfer function proportional to  $\text{Erf}(x/\sigma_x)$ , for simplicity the constant current biased bipolar junction transistor differential pair configuration<sup>4</sup> was employed.

For such a network the transfer characteristic relating the output collector voltage  $y$  and the applied input differential base signal  $x$  is given by

$$y = E_{cc} - \frac{\alpha I_k R}{2} \left[ 1 + \tanh \left( \frac{qx}{2kT} \right) \right] \quad (3)$$

where  $k$  denotes Boltzmann's constant,  $q$  is the charge of an electron,  $\frac{kT}{q} = 26\text{mV}$  at room temperature ( $T = 300^\circ\text{K}$ ),  $I_k$  is the magnitude of the constant current bias source,  $R$  is the collector load resistor,  $E_{cc}$  is the collectors' dc power supply voltage, and  $\alpha$  denotes the transistor "alpha." The shape of this characteristic is governed by the term  $\frac{1}{2} [1 + \tanh(\frac{x}{52\text{mV}})]$ . The remaining terms in Eq. (3) determine the magnitude (peak-to-peak swing) and dc level of the differential pair output voltage. This transfer function has the same general form as the desired error function characteristic. In fact, the reader can show that for the case of  $\sigma_x = 44\text{mV}$  the mean square error between the  $\text{Erf}(x/\sigma_x)$  and its approximating curve is only 2.35% over the range  $-4 \leq \frac{x}{26\text{mV}} \leq 4$ . The differential pair dc output voltage can be eliminated by inserting a dc level shifter or coupling capacitor between the nonlinear network and FM modulator of Fig. 1, or it may be used as a control to adjust the center frequency of the jammer signal generated by the FM modulator.

### C. Experimental Results

In the laboratory a commercially available Gaussian noise generator (GR1382) and an FM modulator (Wavetek 184) were used to implement the system described in Figure 1. The LM3046 integrated circuit transistor array was used for the nonlinear network. Its measured static transfer characteristic was in good agreement with

Equation (3). A four pole maximally flat design Krohn-Hite variable low pass filter was inserted between the Gaussian noise generator and the nonlinear network to control the noise bandwidth and hence the modulation index of the FM modulator output. The pdf of the nonlinear circuit output was monitored with a Polytechnic built "pdf machine."<sup>5</sup> To insure a proper operation of the machine, density measurements were initially made with sinusoids, triangle waves, and square waves and compared with known theoretical responses. Excellent agreement was obtained. The pdf of the GR noise generator was then checked to insure it was Gaussian and zero mean. Typical responses of the differential pair output voltage pdf and the corresponding oscillogram of the FM modulator's output spectrum are shown in Fig. 2(a) and 2(b), respectively. These were measured with the Hewlett-Packard XY plotter and Nelson-Ross spectrum analyzer. The pdf is uniform as expected, and the output signal has a uniform power spectral density, as Woodward's Theorem predicts. The experimental result of Fig. 2(b) reveals that the spectrum is centered around the carrier frequency of 200 kHz, the free running frequency of the FM generator, with a bandwidth of 237kHz. Theoretically, the bandwidth expected is twice the peak frequency deviation which is given by the peak-to-peak swing of the differential pair output voltage ( $I_k R$ ) multiplied by the FM modulator's sensitivity  $K_g$  (kHz/volt). Using the experimental values  $I_k = 1.37\text{mA}$ ,  $R = 2\text{Kohm}$ ,  $K_g = 100\text{kHz/volt}$  yields a theoretical BW of 274kHz. The 14% discrepancy from the measured value can easily be accounted for by component and measurement tolerances and Carson's Rule approximations. Increasing the value of  $R$  or  $I_k$  will increase the peak-to-peak swing at the differential pair output, and hence increase the rms voltage at the FM modulator input and, subsequently, the bandwidth of the modulator's output spectrum. This was verified experimentally in a quantitative manner as the numerical example just presented. This feature provides a useful mechanism for electronically or manually controlling BW. In addition to varying the parameters  $R$  and  $I_k$ , the FM modulation index was varied over the range  $10 \leq \beta_{\text{rms}} \leq 50$  by adjusting the bandwidth ( $f_m$ ) of the Krohn-Hite low pass filter following the GR noise generator. Rectangular output spectrums with Carson's Rule bandwidths were experimentally obtained throughout this range.

National Science Foundation  
ENG 76-244

Army Research Office

DAAG 29-77-G-0232

F. Cassara

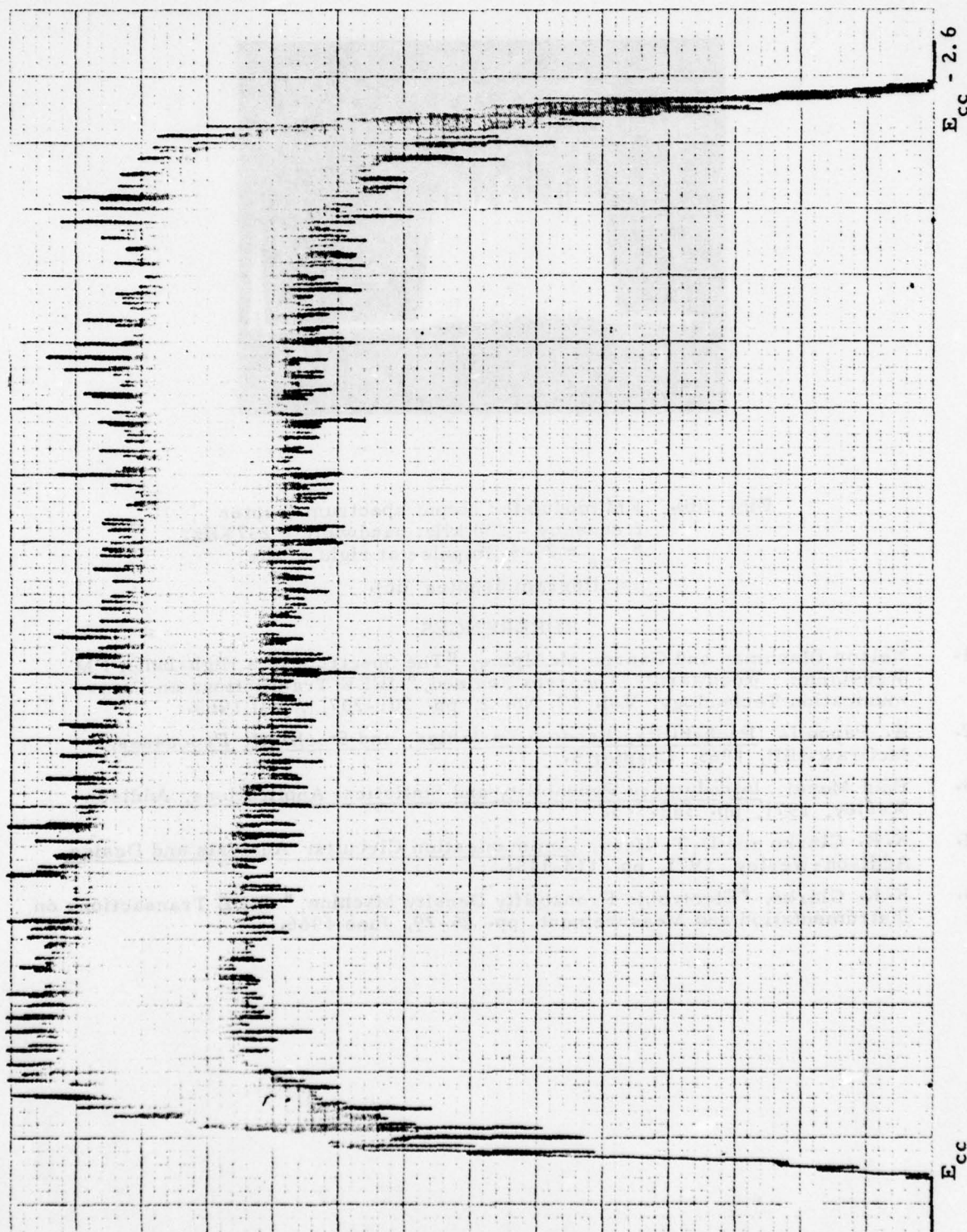


Fig. 2(a). pdf of nonlinear differential pair output with Gaussian noise input ( $I_Q = 1.37 \text{ mA}$ ,  $R = 2 \text{ K ohm}$ ,  $f_m = 4 \text{ kHz}$ ).

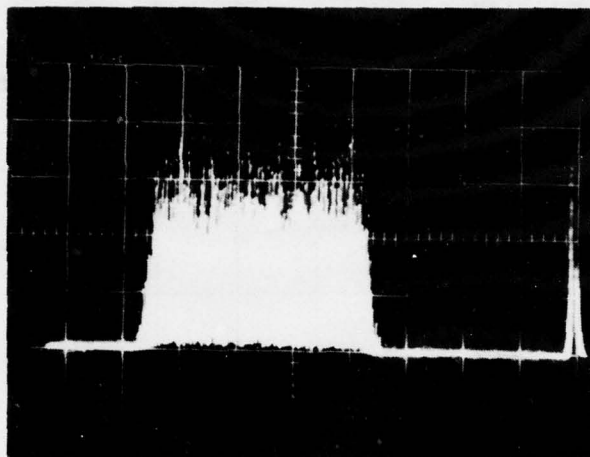


Fig. 2(b). FM modulator output spectrum (center frequency = 200 kHz; bandwidth = 237 kHz;  $\beta_{\text{rms}} = 29.6$ ; impulse at right of the oscillogram denotes dc).

#### REFERENCES

1. Nelson Blachman and George McAlpine, "The Spectrum of a High-Index FM Waveform: Woodward's Theorem Revised," *IEEE Transactions on Communication Technology*, Vol. 17; No. 2, pp. 201-207, April 1969.
2. A. Papoulis, Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1965, Chapter 5.
3. Paul Meyer, Introductory Probability and Statistical Applications, Addison-Wesley, 1971, pp. 342-343.
4. K. K. Clarke and D. T. Hess, Communication Circuits: Analysis and Design, Addison-Wesley, 1971, pp. 114-115.
5. K. K. Clarke, "Electronic Probability Density Machine," *IEEE Transactions on Instrumentation and Measurement*, pp. 25-29, June 1966.

## TRANSIENT ACQUISITION BEHAVIOR OF THE CROSS-COUPLED PHASE-LOCKED LOOP FM DEMODULATOR

F. A. Cassara, H. Schachter and G. Simowitz

In a previous work<sup>1</sup> a novel cross-coupled phase-locked loop (PLL) FM demodulator capable of suppressing co-channel and adjacent channel interferers was derived using maximum-a-posteriori estimation techniques. Experimental results demonstrating such capability even in the presence of strong input gaussian noise, moderate fluctuations in the received signal amplitudes and multiple interferer environments were also presented. The block diagram of this novel FM detector is shown in Figure 1.

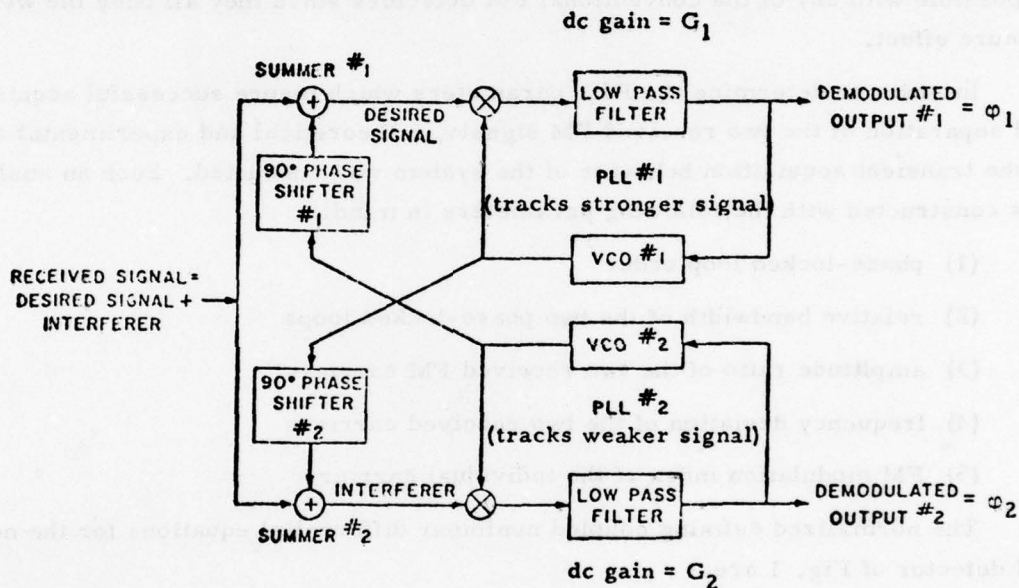


Fig. 1. Cross coupled PLL FM demodulator for suppression of interchannel interference.

The principle of operation can be described as follows: assume the input signal consists of a frequency modulated carrier plus a frequency modulated interferer. Phase-locked loop (PLL) No. 1 locks on to and tracks (by the capture effect) the stronger of the two received FM signals but its voltage controlled oscillator output signal (VCO No. 1) lags by approximately  $90^\circ$ . An additional  $90^\circ$  phase shift is introduced by phase shifter No. 2 so that the signal appearing at phase shifter No. 2's output is  $180^\circ$  out of phase with respect to the stronger received signal. By proper adjustment of the gain constants of summer No. 2, the stronger received signal is

cancelled, leaving only the weaker of the two received FM signals at the input to PLL No. 2. The instantaneous phase of VCO No. 2 output signal tracks the instantaneous phase of the weaker signal but lags by  $90^\circ$ . An additional  $90^\circ$  phase shift is introduced by phase shifter No. 1 producing a signal at the output of phase shifter No. 1 which is  $180^\circ$  out of phase with respect to the weaker of the two received FM signals. The weaker signal can thus be cancelled at summer No. 1 leaving only the stronger signal appearing at the input to PLL No. 1. Since this novel detector has two separate outputs --namely, the outputs of the individual phase-locked loops, it possesses the capability of demodulating both the stronger and the weaker received signals even though they may be co-channel and share the same frequency band. This is a task impossible with any of the conventional FM detectors since they all obey the well known capture effect.

In order to determine the PLL parameters which insure successful acquisition and separation of the two received FM signals, a theoretical and experimental analysis of the transient acquisition behavior of the system was conducted. Such an analysis was constructed with the following parameters in mind:

- (1) phase-locked loop order
- (2) relative bandwidth of the two phase-locked loops
- (3) amplitude ratio of the two received FM carriers
- (4) frequency deviation of the two received carriers
- (5) FM modulation index of the individual carriers

The normalized defining coupled nonlinear differential equations for the novel FM detector of Fig. 1 are:

$$\frac{d\varphi_1}{d\tau} = \alpha_1 [\sin(\psi_1 - \varphi_1) + \eta \sin(\psi_2 - \varphi_1) - \eta \sin(\varphi_2 - \varphi_1)] * h_{L1}(\tau) \quad (1)$$

$$\frac{d\varphi_2}{d\tau} = \alpha_2 [\sin(\psi_1 - \varphi_2) + \eta \sin(\psi_2 - \varphi_2) - \sin(\varphi_1 - \varphi_2)] * h_{L2}(\tau) \quad (2)$$

where

$\psi_1$  and  $\psi_2$  denote the phase modulation of the stronger and weaker received FM carriers, respectively.

$\varphi_1$  and  $\varphi_2$  represent, respectively, the phase modulation of VCO No. 1 and VCO No. 2.

$\alpha_1 = \frac{\omega_{H1}}{\omega_N}$  where  $\omega_{H1}$  denotes the static hold-in range of PLL No. 1 and

$\omega_N$  denotes some selectable "normalization frequency"

$\alpha_2 = \frac{G_2}{G_1} \alpha_1$  where  $G_1$  and  $G_2$  are, respectively, the dc gains of the PLL low pass filters.

$\eta$  denotes the ratio of the weaker to stronger carrier amplitudes.

Digital computer solutions of these equations were obtained for the cases of (a) first order loops with CW interferers, (b) first order loops with a frequency offset CW interferer, (c) second order loops with a frequency off-set interferer, (d) second order loops with a frequency modulated interferer, and (e) second order loops with multipath interference. For each case mentioned the range of loop parameters, i.e., the "stability region" over which the cross-coupled PLL demodulator can successfully separate and demodulate the two received co-channel signals was determined. Such stability regions provide useful design rules. Figure 2 reveals a typical result for a

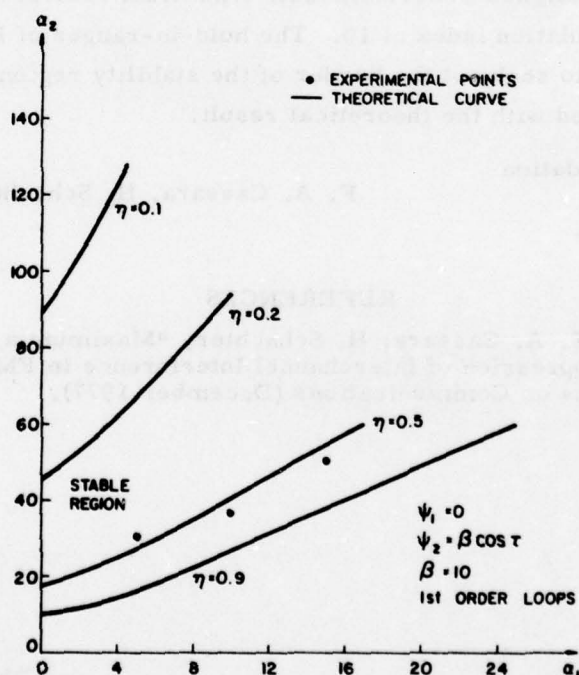


Fig. 2. Transient acquisition stability regions of the cross coupled PLL FM demodulator for first order loops with CW interferer.

first order PLL design with a single CW interferer located at the center of the IF frequency band, i.e., no offset. The weaker signal was frequency modulated by a sinusoid with an FM modulation index equal to 10. The PLL loop parameters  $\alpha_1$  and  $\alpha_2$  shown in Figure 2 were normalized to the modulation frequency,  $\omega_{m2}$ , of this sinusoid, i.e.,  $\omega_N$  was selected to be  $\omega_{m2}$ . For successful separation and demodulation, PLL No. 2's bandwidth,  $\alpha_2$ , must be designed to lie above the solid curves of Figure 2. Since the hold-in-range or equivalently the bandwidth of each PLL is directly proportional to the received signal amplitude, the smaller the value of  $\eta$  the larger we must design  $\alpha_2$  as indicated in the computer solution. For a fixed  $\eta$ , the minimum value of  $\alpha_2$  occurs when  $\alpha_1 = 0$ . Since the hold-in-range of PLL No. 2,  $\eta\alpha_2$  must be at least as large as the peak frequency deviation,  $\beta\omega_{m2}$ , we have  $\alpha_{2,MIN} = \beta\omega_{m2}/\eta$ . The results shown in Fig. 1 reveal  $\alpha_{2,MIN}$  slightly lower than this value. This can be explained by the fact that each PLL's output was filtered on the computer by a four-pole maximally flat designed post detection low pass filter with normalized 3 dB bandwidth = 2 rad/sec. Such a filter reduces the high frequency distortion present in the sharp pulses occurring when the PLL begins to cycle slip. Experimental results have also been obtained to corroborate selected theoretical transient acquisition stability regions. These results are also shown in Fig. 1 for the case of  $\eta = 0.5$ . Such results were obtained on an experimental model designed to accommodate sinusoidal carriers at frequencies of 200 kHz with FM modulation index of 10. The hold-in-ranges of PLL No. 1 and No. 2 and  $\omega_{m2}$  were varied to seek out the border of the stability regions. Reasonably good agreement was obtained with the theoretical result.

National Science Foundation  
ENG 76-244  
Army Research Office  
DAAG 29-77-G-0232

F. A. Cassara, H. Schachter and G. Simowitz

#### REFERENCES

1. T. S. Sundresh, F. A. Cassara, H. Schachter, "Maximum a Posteriori Estimator for Suppression of Interchannel Interference in FM Receivers," IEEE Transactions on Communications (December 1977).

## SECOND ORDER GREEDY ALGORITHMS FOR CENTRALIZED TELEPROCESSING NETWORK DESIGN

A. Kershenbaum and R. Boorstyn

A. Introduction

The problem considered is that of finding an optimal (minimum cost) design for a centralized telecommunication network given a set of terminal locations, traffic magnitudes between these locations, and a single common source or destination (central site). In order to retain simplicity and low cost in the terminal hardware, such networks are configured as trees comprised of communication facilities of a single capacity. Thus, the optimal solution to this problem is a capacitated minimal spanning tree (CMST); i.e., a tree of minimum total length satisfying a constraint or set of constraints that limit the total traffic and/or number of nodes in any subtree rooted at the central site. In centralized networks, such subtrees are called multipoint lines. This is a continuation of work previously reported on.

More formally, we are given a set of locations (nodes),  $A = \{a_i | i=0, 1, \dots, n\}$ , where  $a_0$  is the central site; a symmetric distance measure,  $D = \{d_{ij} | i, j=0, 1, \dots, n\}$ , giving the cost between any pair of locations; and a constraint,  $M$ , on the total number of nodes or traffic in a multipoint line. We seek a spanning tree,  $T$ , rooted at  $a_0$ , satisfying the constraint, and of minimum total length.

This formulation is quite general. All we require of the cost function is that the cost of a link,  $d_{ij}$ , not depend upon what other links are present in the solution. The constraint is, likewise, quite general. One can, in fact have more than one constraint and the constraints may be on total traffic, number of nodes, nodal degree, number of links in cascade, or any other quantity associated with the design. We require only that if some subset,  $S$ , of  $A$ , does not satisfy the constraints then no other subset,  $S'$ , containing  $S$ , satisfies the constraints. We also assume that a star solution; i.e., all nodes connected directly to the center, is feasible. This can always be made so by splitting a location into two or more nodes.

Several heuristics have been developed for the solution of this problem and optimal techniques exist as well. The optimal techniques<sup>1, 2, 3</sup> are branch and bound procedures which, in general have running times which are exponential in the number of nodes. While they are, thus, not practical as design procedures, results obtained using them, in particular a recently developed procedure,<sup>7</sup> have led to some insight into the refinement of heuristic procedures. The procedure described below is an outgrowth of this work.

Currently, the most widely used procedures for the solution of the CMST problem are heuristics<sup>3,4,8,10</sup> which produce solutions within 5% of the optimum (in cases where the optimum is known) and which have running times which are a low order polynomial, generally between quadratic and cubic, in the number of nodes. Karnaugh,<sup>6</sup> has developed a family of Second Order Greedy Algorithms (SOGA's) which iterate the above heuristics (which he refers to as F (First) OGA's, and have longer running times but produce results generally 2 to 3% better than the above heuristics. The procedures described below are variants of the general SOGA procedures described by Karnaugh and are shown to be considerably faster than his; indeed, in many cases they are competitive with the simpler procedures previously used.

### B. Procedural Description

The most often used heuristic solutions to the CMST problem share the following properties: (1) their running time is a polynomial (usually of order between 2 and 3); (2) in the absence of constraints they will yield a minimum spanning tree (MST); and (3) the quality of the solution (i. e., the amount by which it differs from the optimum) is not controllable and, except very loosely, is not known.

The basic heuristics used to solve CMST problems can be divided into two categories - primal procedures which seek to improve a feasible starting solution or partial solution and dual procedures which seek to make a low cost (infeasible) starting solution feasible. We concentrate on primal procedures, having found them to be generally more flexible. It has been shown<sup>8</sup> that most such procedures fall within the framework of the following scheme:

- (1) Start with each node on a separate multipoint line directly connected to the center. Associate a weight,  $w_i$ , with each node,  $i$ , by applying a given rule (w-rule). For each potential link  $L_{ij}$  interconnecting a pair of nodes,  $i$  and  $j$ , define a trade-off function,  $t_{ij}$ , as  $d_{ij} - w_i$ .
- (2) Consider the (not previously considered)  $L_{ij}$ , for which  $t_{ij}$  is minimum. If nodes  $i$  and  $j$  are in separate multipoint lines and merging these links does not violate any constraint, then add  $L_{ij}$  to the network, replacing  $L_{i0}$  or  $L_{j0}$ , whichever is more costly. If not, reject  $L_{ij}$ .
- (3) Update the  $w_i$  and  $t_{ij}$  and return to Step 2 until no further gain can be obtained. Usually,  $w_i = w_j$  for nodes  $i$  and  $j$  in the same multipoint line.

The implementation of such a procedure has been considered in detail.<sup>9</sup> A careful implementation is shown to be of order  $N^2 \log_2 N$  if one wishes to consider all branches and of order  $NK \log_2 N$  if one only examines the  $K$  nearest neighbors of each node and if the w-rule itself is not too computationally complex. In practice, the most widely used w-rule is to set  $w_i = d_{i0}$ ; i. e., the weight of each node in component

$i$  is set to be the minimum distance between a node in the component and the center. This is known as the Esau-Williams algorithm. Karnaugh<sup>6</sup> described an alternate implementation of the Esau-Williams procedure which has a different computational complexity, generally between quadratic and cubic, but which produces identical results. While the SOGA's described below can work with any imbedded FOGA, for the sake of comparison with previous work we performed experiments with an imbedded Esau-Williams algorithm.

In order to implement a SOGA, one must decide which arcs one will attempt to force in or out of the solution. Karnaugh suggested two possibilities:

- (1) Inhibit - At each stage in its execution, the FOGA brings in one arc connecting two previously unconnected sets of nodes. The Inhibit loop successively prevents each of these mergers from taking place. Thus, one iteration of the Inhibit loop involves  $N$  (where  $N$  is the number of terminals) iterations of the FOGA each of which prevents a single pair of node clusters from merging. At the end of each iteration of the Inhibit loop the best of the  $N$  generated solutions is kept and used in place of the original (uninhibited) FOGA solution and the Inhibit loop is repeated until no further progress is made.
- (2) Join - For each node  $a_i$  find its nearest neighbor,  $b_i$ , and nearest neighbor which is also closer to the center than  $a_i$ ,  $c_i$ . Successively, one arc at a time, force arcs  $(a_i, b_i)$  and  $(a_i, c_i)$  into the solution if they are not already present. One Join loop consists of a sequence of executions of the FOGA, each with a single forced arc. The Join loop is iterated, starting from the best solution obtained during the previous loop, until no further progress is made. The number of FOGA iterations in a Join loop is bounded from above by  $2N-1$  but is, on the average somewhat less than  $N$ , the exact number being a function of the particular problem.

Thus, both the Inhibit and Join procedures have running times of order  $CN$  times the running time of the embedded FOGA, the Join procedure being somewhat faster. The factor  $C$  is the number of iterations required for convergence, i.e., to reach the point of no further progress. It would be possible to greatly improve the running time of the SOGA if one could restrict one's attention to a small (i.e.,  $\ll N$ ), subset of the arcs. Also Karnaugh mentioned that the Join procedure, while somewhat faster than Inhibit, is weaker because it restricts itself to local transformations involving nearest neighbors. Using the techniques in Ref. 7 it was possible to obtain optimal solutions for a class of interesting problems and make observations about characteristics of optimal solutions. This led to the characterization of arc subsets which are small and at the same time reasonably effective for use as candidates in a SOGA. For a CMST problem with 18 nodes and a constraint on the size of any subtree of  $m=3$ , Figure 1a is the solution yielded by the Esau-Williams algorithm. Figure 1b is the unconstrained MST on the same problem. Figure 1c is the optimal solution. Note that the two arcs present in the optimal solution but not present in the Esau-Williams solution are  $(G, O)$  and

(B, K) and are MST arcs. While it is not necessarily so that all arcs present in the optimal solution and not in the Esau-Williams solution are MST arcs, in all the experiments run and checked to this end, invariably at least one arc in the optimum but not in the Esau-Williams solution was an MST arc.

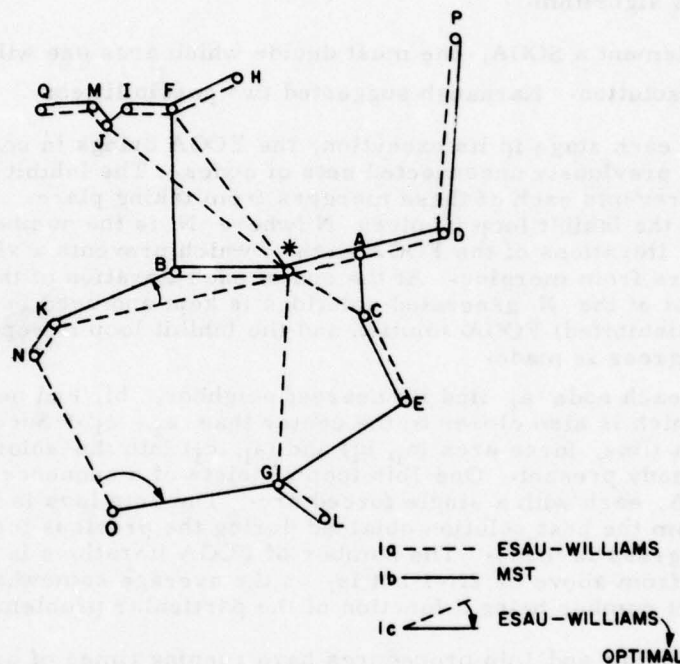


Fig. 1. Comparison between Esau-Williams' and optimal solution.

This is not to say that it is conjectured that one must always be able to find an optimum solution which contains an MST arc not part of the Esau-Williams solution (when the Esau-Williams solution itself is not optimal). It merely strengthens the point that MST arcs are reasonable arcs to examine when seeking to improve a heuristic solution.

Even without this evidence, however, it is intuitively appealing to consider such arcs. Most primal heuristics share the property that they proceed by connecting nodes to their nearest feasible neighbors.<sup>8</sup> The algorithms differ only in the order in which they consider the nodes. Thus, the solutions differ only in that the distance to

a node's nearest feasible neighbor may differ depending upon when the node is considered during the execution of the algorithm. If such a heuristic fails to find the optimum, it will be because the algorithm erred by waiting too long before consideration of some critical node (or nodes) and hence missed being able to include some desirable arc (or arcs) in the solution. It is quite likely that one or more of these arcs will be MST arcs; indeed, as mentioned above, this was always the case with the experiments carried out.

Thus, a heuristic is proposed which attempts to improve solutions generated by some primal heuristic by forcing the inclusion of one or more MST arcs which the primal heuristic left out. This gives rise to the following procedure:

**Step 1:** Generate a MST,  $T_m = \{a_j | j=1, \dots, n-1\}$ . Generate a feasible tree,  $T_f = \{b_j | j=1, \dots, n-1\}$ . Find  $S = T_m - T_m \cap T_f$ . (For problems of interest,  $S$  is nonempty.)

**Step 2:** For each subset  $S_1 \subset S$ : set  $S_2 = S - S_1$ . Remove all  $a_j \in S_2$  from the network. Start a solution by including all  $a_j \in S_1$ . Apply the primal heuristic to complete the generation of a solution.

It should be noted that an alternative procedure would be to include  $a_j$  in  $S_1$  but not to exclude  $b_j$  in  $S_2$ . It was felt, however, that if elements in  $S_2$  were retained as candidates for inclusion, duplicate solutions would be likely to result. The exclusion of elements in  $S_2$  guarantees unique solutions and hence should increase the likelihood of generating improvements.

### C. Computational Experience

In order to compare the effectiveness and efficiency of the procedural scheme described in the previous section, a series of experiments was run. Nodes were generated with random locations over a unit square and unit traffic; problems with  $N = 40, 60, 90$ , and  $120$  nodes were generated and constraints varying between  $N/20$  and  $N/4$  were examined.

Alternatively, a constraint on traffic could have been used but we did not feel that significantly different results would have been obtained. Our first objective was to examine the running time of our implementation of the Esau-Williams Algorithm within the context of the Unified Algorithm.<sup>8</sup>

Karnaugh<sup>6</sup> described an alternate implementation of the Esau-Williams procedure and reports on its running time on an IBM 370/158. These times are shown in Table I together with those obtained using the above implementation of the Esau-Williams procedure on a set of problems similar to those reported on in Karnaugh's paper; the runs were made on a DEC KL20/50 processor, which is somewhat slower

than a 370/158. As can be seen, there is a considerable difference between the running times of the two procedures, the one described here being faster. While it is somewhat difficult to compare the running times of the procedures based on such a small number of sample runs, especially since the computers and compilers are different, the difference in running times is great enough to warrant the conclusion that the Unified Algorithm is faster and of a lower order of complexity.

Table I. Esau-Williams Algorithm  
Runtime - (Initial execution)

Number of Nodes	Runtime (seconds)	
	Karnaugh	Unified
40	3.03	.51
60	3.25	1.11
90	4.17	2.63
120	5.64	4.67

In general, the second and subsequent calls to the FOGA do not require as much running time as the initial call since the initialization of the problem, i.e., finding nearest neighbors and setting up the trade-off functions, does not have to be repeated. Also, one can start later FOGA iterations with a partial solution obtained in the previous call. Specifically, if one is up to the  $K^{\text{th}}$  call to the FOGA within a loop, one can bring in the same arcs brought in on the previous FOGA call until the new forced inclusion or exclusion induces a change in the solution. Table II (in a manner similar to Table I) shows a comparison of the running times for the second and later calls to the Esau-Williams procedure implemented both ways. Again we see the Unified Algorithm is faster.

Table II. Esau-Williams Algorithm  
Runtime - (Later Executions)

Number of Nodes	Runtime (seconds)	
	Karnaugh	Unified
40	.07	.03
60	.16	.06
90	.32	.12
120	.57	.19

Next, we examined the size of the subsets,  $S$ , generated, i.e., the average number of MST links not present in the FOGA solution. Most primal heuristics generate subtrees which are MST's on the set of nodes they contain along with the center. Most of the arcs in these subtrees (with the exception of the arc directly connected to the center) will be MST arcs. Thus, one might expect that  $|S| \approx \frac{n}{m}$ , and a vast improvement can be made over a blind branch exchange or SOGA procedure. However, this turns out to be slightly conservative. In practice, for  $n \leq 20$  and  $v_0$  in the geographic center, and for  $n \leq 15$  and  $v_0$  in the corner,  $|S| \approx \frac{n}{m}$ . For larger networks,  $|S|$  ranges between  $\frac{n}{m}$  and  $\frac{n \log n}{m}$ , where  $|S|$  for networks with non-centered roots is generally larger than for the same network with  $v_0$  in the center.

There are, in fact,  $2^{|S|}$  subsets of  $S$ . Thus, for moderately large tightly constrained problems,  $S$  could grow large enough to make evaluating all subsets impractical. Furthermore, it is reasonable to assume that not all arcs in  $S$  interact with one another; i.e., improvements in separate parts of the network can be found and justified independently of one another. Further, in running the experiments, it was found that if a group of arcs yielded an improvement when forced into the solution, subsets of these arcs often yielded improvements too. Thus, the heuristic was modified by only considering subsets  $S_1 \subset S$  such that  $|S_1| \leq K$  for some given  $K$ . The best subset,  $S_1^*$ , is found and permanently forced into the solution. We then set  $S = S - S_1^*$  and repeat the procedure until no further improvement can be made. As multiple branch exchanges are often required to effect improvements, the value of  $K$  chosen should be large enough to ensure that most advantageous exchanges will be found and at the same time small enough to keep the procedure computationally tractable;  $K=2$  seemed to work well. Experiments were run with larger values of  $K$ ; only in isolated cases was any improvement over  $K=2$  obtained; in no case was an improvement in excess of 1% observed. Thus, it was decided in all the remaining experiments to use the extended procedure (which proceeds from the best solution found) and restrict the examination to subsets of cardinality less than or equal to two.

Using this modification of the new heuristic on a large number of problems, it was observed that in general, forcing arcs between nodes which are close to the center seemed to have the greatest effect on the value of the solution. This was probably a consequence of the fact that the Esau-Williams algorithm starts with nodes far from the center and hence, dealing with nodes near the center first radically changes the solution value.

Large subsets, in some cases all of  $S$ , sometimes yielded the best solutions. This means that one cannot simply consider subsets of limited size without iterating to subsequently force additional arcs into the solution.

In particular, it was found that for small networks ( $n \leq 20$ ) with centered  $v_0$ , there was little difference (usually  $\leq 1\%$ ) in performance between the new heuristic and the Esau-Williams solution. This is almost certainly because both procedures were generating near-optimal solution. However, for larger networks (up to  $n = 100$ ) and centered  $v_0$ , particularly for tightly constrained problems, the new heuristic performed noticeably better with improvements averaging about 1.5%. Figure 2 shows the relative improvement obtained using the new heuristic relative to the Esau-Williams algorithm. Each point on the curve represents an average over 12 sample runs. As can be seen, the improvement increases as the problem size does. Significant savings are obtainable for realistic problems.

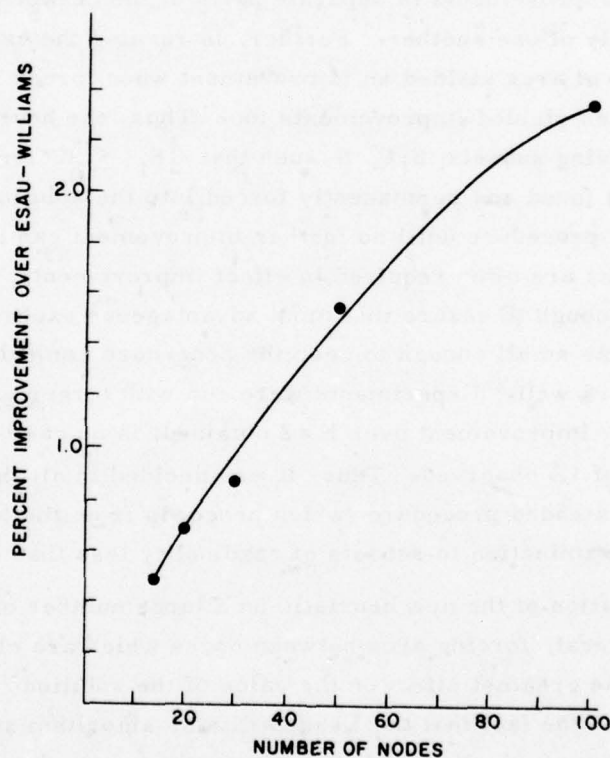


Fig. 2. Improvements obtained with new heuristic.

Surprisingly, the improvement was only weakly correlated with problem tightness (ratio of  $m$  to  $n$ ); the new heuristic achieved slightly larger savings in more tightly constrained problems. Related to this, is that substantial savings were achieved even for very loosely constrained problems. In such cases, the constrained solutions differed only slightly from the MST, as expected.

Results were even more encouraging for large networks with  $v_0$  in the corner, such networks may be viewed as one-fourth of a network of  $4n$  nodes with  $v_0$  in the center. For such problems, improvements averaging 4% and as high as 8% were observed. Thus, the procedure appears to be of value for many realistically-sized problems with tight constraints.

A straightforward implementation of the Esau-Williams procedure has a computational complexity of order  $n^2 \log_2 n$ . A more careful implementation<sup>9</sup> can reduce the complexity to order  $n \log_2 n$ . As discussed previously,  $S$  was found to range between  $\frac{n}{m}$  and  $\frac{n \log n}{m}$ . Since we examine subsets of cardinality at most 2, subsets are examined on each iteration. At worst,  $\frac{|S|}{2}$  successive subsets of cardinality could be introduced and thus, the procedure could iterate at most  $\frac{|S|}{2}$  times. In practice, the number of iterations grows more slowly than  $|S|$ .

Thus, a careful implementation of the procedure has a complexity of  $|S|^3 n \log n$ . If, for instance,  $|S| \approx n^{1/2}$ , the procedure's complexity is  $n^{5/2} \log n$ .

Finally, a version of the Inhibit and Join procedures (as described in Ref. 6 but using the Unified Algorithm in place of the alternate implementation of Esau-Williams) were coded up and run on another set of problems and the heuristic described in this paper was then run on the same set of problems. Thus, we were able to directly compare the quality of the obtained solutions as well as the running times of the new heuristic with Inhibit and Join. The running times are summarized in Figure 3. As can be seen, the new heuristic is much faster than Inhibit, and faster than Join. Indeed, it is only 2-3 times slower than the Unified Algorithm.

The quality of the solutions obtained varied. In some cases the new heuristic outperformed Inhibit; in others Inhibit performed better. On the whole, the quality of the solutions obtained with Inhibit were 2.6% better than those obtained using the Esau-Williams procedure, while those obtained using the new heuristic were 1.9% better than Esau-Williams. Both Inhibit and the new heuristic outperformed Join fairly consistently; Join averaged a 1.5% improvement over Esau-Williams.

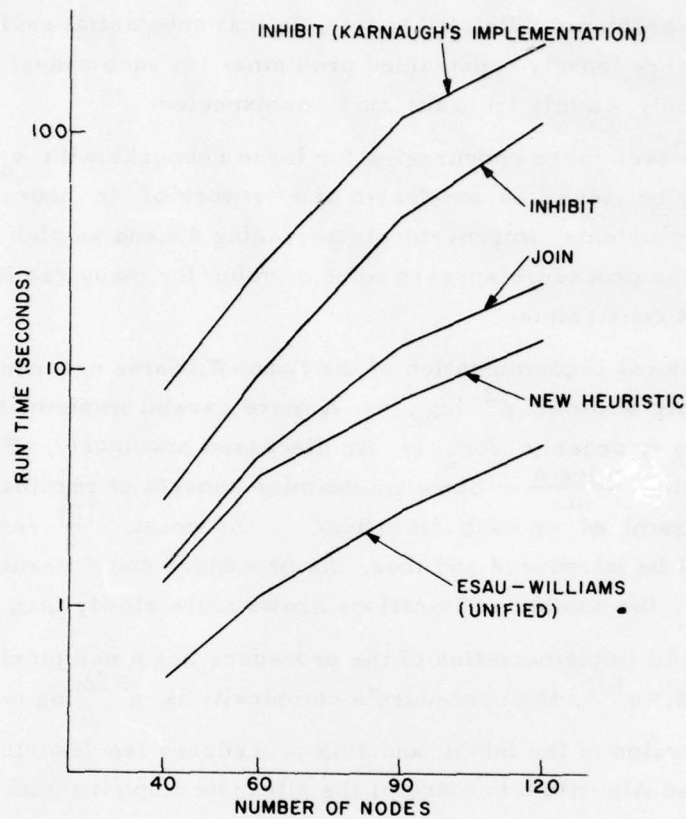


Fig. 3. Comparison of runtimes.

#### D. Conclusion

We found the new heuristic to be of definite value as it obtained solutions roughly 2% better than the Esau-Williams procedure without greatly increasing the running time. In comparison with Karnough's SOGA's the new heuristic is much faster and obtains solutions somewhat better than Join and somewhat worse than Inhibit. We conclude that the new heuristic is useful as a practical design procedure even when imbedded in a larger procedure which solves more global problems.

A. Kershenbaum and R. Boerstyn

## REFERENCES

1. Chandy, K. M. and R. A. Russell, "The Design of Multipoint Linkages in a Teleprocessing Tree Network," IEEE Trans. on Computers, Vol. C-21, pp. 1062-1066 (1972).
2. Chandy, K. M. and T. Lo, "The Capacitated Minimum Spanning Tree," Networks, Vol. 3, No. 2, pp. 173-182 (1973).
3. Elias, D. and M. J. Ferguson, "Topological Design of Multi-Teleprocessing Networks," IEEE Trans. on Communication; Vol. C-22, pp. 1753-1761 (1974).
4. Esau, L. R. and K. C. Williams, "On Teleprocessing System Design, Part 11," IBM Syst. J., Vol. 5, No. 3, pp. 142-147 (1966).
5. Frank, H., I. T. Frisch, W. Chow and R. Van Slyke, "Optimal Design of Centralized Computer Networks," Networks, Vol. 1, pp. 43-57 (1971).
6. Karnaugh, M., "A New Class of Algorithms for Multipoint Network Optimization," IEEE Trans. on Communication, Vol. C-24, No. 5, pp. 500-505 (May 1976).
7. Kershenbaum, A. and R. Boorstyn, "Centralized Teleprocessing Network Design," Proceedings of the NTC (December 1975).
8. Kershenbaum, A. and W. Chou, "A Unified Algorithm for Designing Multidrop Teleprocessing Networks," IEEE Trans. on Communications, Vol. C-22, No. 11, pp. 1762-1772 (1974).
9. Kershenbaum, A., "Computing Capacitated Minimal Spanning Trees Efficiently," Networks, Vol. 4, No. 4 (1974).
10. Reinfield, N. V. and W. R. Vogel, Mathematical Programming, Prentice-Hall, Englewood Cliffs, N. J. (1958).

## ADAPTIVE ROUTING IN NETWORKS

R.R. Boorstyn and A. Livne

Most adaptive routing schemes perform about as well as nonadaptive schemes, when evaluated in a fixed environment.<sup>1</sup> However, they do adapt to changes in network topology and input statistics. The reason they do not show a significant improvement is that they are actually quasi-static, in that they sense and respond to the above changes slowly. Their goal is to select good paths.<sup>2</sup>

We focus our attention on the node, which may be viewed as a multiple server queuing system. Most adaptive routing schemes operate the node as a collection of single server queues. If the node is operated as a multiple server queue, an advantage at the node of a factor equal to as much as the number of servers (outgoing branches) can be obtained. However, the control over good paths may be lost. We show that we can simultaneously obtain both good paths and improved node performance. Furthermore, the approach lends itself to analysis. We also consider limits and extensions.

Many computer communication networks use dynamic routing schemes to compensate for input traffic variations, to respond to changes in topology, and to take advantage of temporary changes in loading in different paths. These adaptive schemes are complex and are usually chosen and verified by extensive simulations. Invariably during the design of a network they are replaced by analytically tractable non-dynamic (static) schemes. We present here a dynamic routing scheme for which we have been able to derive approximate analytical models. Furthermore, we can establish the efficiency of this scheme, especially in heavily loaded situations.

A typical static routing scheme would operate as follows. Consider as separate commodities the messages originating at a particular node and destined for a second node in the network. There are usually several good paths connecting these nodes. The static routing scheme would specify the optimum proportion of traffic to be routed over each path. Efficient algorithms exist for design of this type of routing.<sup>3</sup>

We can identify one particular problem with this approach. Although good paths are indeed found, any node essentially operates as a collection of single server queues -- one queue for each outgoing branch. Considering the node as a queue with several potential servers this is not an efficient manner of operation. Indeed if a node had  $k$  outgoing branches and was operated as a queue with  $k$  servers then when heavily loaded the time delay would be reduced by a factor of  $k$ . Conversely, the throughput can be increased.

If the node was operated as suggested above, messages would wander aimlessly through the network and the total performance would be abysmal. Our approach is to

retain the good paths for commodities and yet still get the benefit of the faster performance at the node.

Briefly our scheme is as follows.<sup>4</sup> Consider a node as a single queue with several servers (output channels). For a particular commodity, i.e., a message with a certain destination, the use of some of these servers would cause the messages to be sent along "bad" paths -- either too long or too heavily loaded. Thus for each commodity and at each node we specify a subset of the output channels as allowable and permit the message to use any allowable channel according to some discipline. Each commodity appearing at the node has its own set of allowable channels. These restrictions force messages to use "good" paths.

The assignment of allowable branches at each node for each commodity is one level of our adaptive routing scheme. These assignments are based on essentially global information of topology, flows, and long term averages, and may be adaptive in a quasi-static way. The second, and truly dynamic, level of our routing strategy is local and involves the queue discipline at each node. As an example, consider a node with two outgoing channels (servers). All message commodities fall into three classes. Two of these must use only one of the servers and have no choice. Messages join the appropriate queue and are served in turn. The third class of messages may use either of the two servers and join a third queue.

The best strategy is to give priority to the messages that have no choice, i.e., are dedicated to one of the servers. Another strategy is to allow messages in the third category (non-dedicated) to join the shorter of the two dedicated queues. We have evaluated the performance of both strategies, and other similar strategies, for this simple two server model, and for more complex models. The measure of performance was the average time delay for all messages traversing the node.

Although individual commodities perform very differently, the average delay for all is fairly insensitive to the different disciplines we have investigated. (Of course, other less appropriate disciplines will behave poorly.) Essentially so long as a relatively modest amount of the traffic has choice, and for a moderately high utilization, the performance of the node is close to that of a multiple server queue. This has been observed for a wide variety of combinations of servers, commodities, and flows.

It has been shown that in the limiting case of heavy traffic, using diffusion techniques, that a join-the-shortest queue discipline converges to the performance of a multiple server queue so long as there is some traffic that has choice that spans, in some combination, all the servers.<sup>5,6</sup> This property is also true for a wide variety of configurations.

We have evaluated this dynamic routing strategy using analytic techniques for simple configurations, approximate techniques for more complex models, simulation, numerical analysis, and limiting arguments in heavy traffic. All results reflect a robustness in performance as described.

In general, we have found, that the node retains its performance advantage as a multiple server queue as long as a modest amount of the traffic has choice. We have developed approximate and fairly accurate analytical models to calculate the performance of these nodes.

We have also developed analytical methods to imbed these nodes in a network and to calculate the overall performance of the network. Basically, we find that if the average number of output channels per node is  $k$ , then the time delay for messages in a heavily loaded network can be reduced by as much as a factor of  $k$  when this dynamic routing is used.

We have been studying an improvement in the routing strategy by allowing the service discipline to depend not only upon queue lengths at the local node but on the status of neighboring nodes. Simple canonical models exhibiting this behavior can be analyzed in the heavy traffic limit using diffusion techniques.

This report is an updated version of a previous report and was presented at the AFOSR Workshop on Communications Systems and Applications, Provincetown, Mass., September 1978.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

R.R. Boorstyn and A. Livne

National Science Foundation  
ENG73-04232

#### REFERENCES

1. M. Gerla, "Deterministic and Adaptive Routing Policies in Packet-Switched Computer Networks," Proc. 3rd Data Communications Symp., St. Petersburg, Fla., pp. 23-28 (November 1973).
2. R.G. Gallager, "A Minimum Delay Routing Algorithm Using Distributed Computation," IEEE Trans. Communications, Vol. COM-25, pp. 73-85 (January 1977).
3. D.G. Cantor and M. Gerla, "Optimal Routing in a Packet-Switched Computer Network," IEEE Trans. Computers, Vol. C-23, No. 10, pp. 1062-1068 (October 1974).
4. A. Livne and R.R. Boorstyn, "On a Technique for Dynamic Routing," Proc. Natl. Telecommunications Conf., Dallas, Texas, pp. 42.2-1; also, "Dynamic Routing in Computer Communication Networks," A. Livne, Ph.D. Dissertation, Polytech. Inst. of New York (July 1976).
5. G.J. Foschini and J. Salz, "A Basic Dynamic Routing Problem and Diffusion," IEEE Trans. Communications, Vol. COM-26, No. 3, pp. 320-327 (March 1978).
6. G.J. Foschini, "On Heavy Traffic Diffusion Analysis and Dynamic Routing in Packet Switched Networks," in Computer Performance, Chandy and Reiser, eds., North Holland Pub. Co., pp. 499-513 (1977).

## MODELING OF DIGITAL SYSTEMS USING MICROPROCESSORS

D.R. Kaufman and E.J. Smith

A. Introduction

Simulators for purposes of logic and design verification play a major role in any design automation system. This report presents a logic simulator consisting of a network of microprocessors modeling the subunits of a digital system. The simulator is highly interactive, has a fast response time, and is capable of modeling a large digital system. An analysis of the advantages and disadvantages of the two commonly used methods, breadboarding and software simulation, is presented. To overcome some of their limitations the new modeling system is proposed.

Logic verification through breadboarding is used for small and medium size systems. The prototype provides an accurate model of the logic but not of the delay characteristics of the device, since the former is usually built using different technology components. One of the advantages of a prototype is that it can be connected to other systems; thus interface problems can be detected. When a subunit is finally built the corresponding hardware from the model can be replaced and thus the prototype can act as a testing tool.

Breadboarding allows for high interaction between the designer and the prototype. The debugging tools (oscilloscopes, logic analyzers) are familiar to the designer and thus easy to use. One of the disadvantages of breadboarding is initialization of the logic subunit, since the prototype can power-up in different states and thus it is difficult to know whether the input stimuli have deterministically driven the prototype to a specified state. For any large digital system it becomes very difficult and expensive to build a model, since the components used usually contain a small number of gates. Another drawback is that the prototype hardware cannot be used in a different project.

Software simulation makes use of an available general logic simulator that provides the user with a large number of commands. The program accurately models both the logic and timing characteristics of the device. There is no initialization problem since every gate is assumed to be in an undefined state at the beginning. The simulator is typically a large software system shared by more than one user. Also, due to the general nature of the program there is a great deal of overhead processing not necessary for simple logic verification. The drawbacks of software logic verification are poor turn-around times, especially for large digital systems, and low interaction between the user and the simulator.

The simulator that will be considered in the present work was designed to overcome some of the limitations for the other modeling systems. The system of an array

of microprocessors operating under centralized control. Each microprocessor card models a subunit of the system to be simulated. Modeling of the overall is done by software using a standard simulation algorithm for logic verification. The input and output pins of the card are functionally identical to the subunit pins. By wiring the cards together we obtain the model of the whole system. The system is very flexible since it allows for mixed models to be used, i.e., some cards can contain register-transfer-level descriptions while others can contain gate-level descriptions. The possibility exists for a microprocessor card to be replaced by "real" hardware thus allowing the system to be used as a testing tool also. This approach has all the advantages of breadboarding and solves most of the problems. Initialization of the logic circuit is no longer a drawback since simulation is done in software. Building a system model is no longer a major effort since the tasks to be done are to provide the microprocessor card with a description of the logic circuit and to wire the cards together. One major advantage of the system is that the hardware is reusable in other projects. The system can be connected to existing digital systems, but with some limitations due to the fact that the microprocessor array is slower than a hardware prototype. The system has advantages over software simulation in that it is highly interactive and the response time is much better because each subunit is simulated by a separate microprocessor. A data structure for representing a logic net and an algorithm for nominal delay simulation will be presented in a form suited for implementation on a microprocessor. We note here that the main objectives are to provide an accurate model, fast response to input stimuli, and capability to handle large digital units. The architecture of such a system will be presented and the basic components described in detail. Attention will be paid to minimizing the amount of logic necessary to control the array of microprocessors and basic operating procedures will be discussed.

#### B. Simulation Data Base

In this section a data structure for the representation of logic circuits is presented. We start by defining some of the terms used and the features of the simulator. One of the goals in determining the format of the data base is to provide the minimum amount of information to the model unit in order to save memory and model set-up effort. One of the basic assumptions is that an 8-bit microprocessor will be used in the model unit to simulate the logic net.

Each model unit has a set of inputs functionally equivalent to the inputs of the subunit being simulated. These inputs will be called primary inputs (PI) and similarly the outputs will be called primary outputs (PO).

For each gate (sometimes referred to as node) the following information is needed:

the logic function (FCN), the propagation delay (DLY), the set of inputs (INP), and the gate output (CUT). In order to describe the logic unit we need a table that gives the fan-out list of gates (sometimes referred to as sinks (SNK)) for each node. The list also indicates the input pin (PIN) fed by the node. In order to speed up the retrieval of a node's sinks we associate with each node a pointer (PNT) that indicates the beginning of the fan-out list in the sink table. The end of the record is indicated by a bit which is set to 1 to indicate the last entry.

The information needed by the simulator is divided into two groups: static and dynamic. The static information is not altered during simulation and can therefore be stored in Read-Only Memory. Also, the static information is the minimum required to represent a logic network all other information being derived from it. The static information consists of logic function, propagation delay, and fan-out list. The dynamic information is stored in Random Access Memory and is derived by the microprocessor from the static tables during initialization or altered during simulation.

Besides the variables discussed so far, we need two flags which are used by the simulator. They are: the initialization flag (INI) used to initialize gate inputs, and a storage flag (STO) used during simulation. The use of these flags will be explained in the next section describing the simulation algorithm. In order to show an example of coding of all the above variables we must define the features of the simulator.

The model unit can simulate a logic net of up to 2,000 gate elements. The range of propagation delays is from  $\emptyset$  (for wired logic) to a maximum of 15 units. The number of logic functions that can be simulated is 15 and each function is represented by an integer in the range of 0 to 14. For example: PI(0), PO(1), AND (2), OR (3), ..., etc. A gate is limited to a fan-in of 16 inputs while the fan-out is unlimited. Figure 1 shows the format of the simulation data base.

The next data structure to be discussed is the implementation of the time-flow mechanism. The time element will be represented by an array of queues where each queue corresponds to a time unit. The number of entries in the queue-array is given by the maximum propagation delay. Each queue stores the nodes that were scheduled to be processed at that point in time. The current queue to be processed is indicated by a clock variable (CLK) which is incremented modulo  $(d + 1)$  where  $d$  is the maximum gate propagation delay. Since each queue has a variable length it is necessary to employ some memory management technique. We chose to use dynamic memory allocation and implement each queue by a linked list. Initially all the memory allocated for the time-queues is linked together in a list called the available list and is pointed to be the available pointer (AVL). In order to operate on the queue (insert and delete)

we need a start-of-list pointer (STR) and an end-of-list pointer (END). These pointers are stored in the time array.

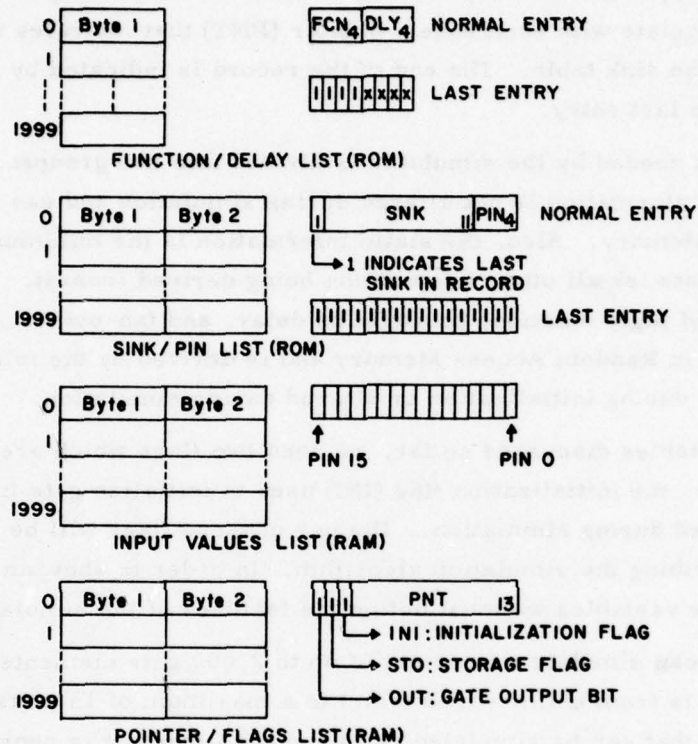


Fig. 1. Simulation data base format.

When processing the nodes in a queue the simulator needs a stack to temporarily store the nodes' sinks. The stack is operated on through the use of a stack pointer (STK) for deleting and inserting elements. The reason behind the use of a stack and the stack operations will be described in the next section. Figure 2 shows the implementation of the time-queues and of the stack.

The memory requirements for the data base are as follows: Using an average number of sinks = 2 we need 8K for the sink table. We allocate 8K for the time queues and the stack. The rest of the requirements are obvious from the formats of the tables. Summarizing, we need 10K of ROM and 16K of RAM to handle 2,000 logic gates.

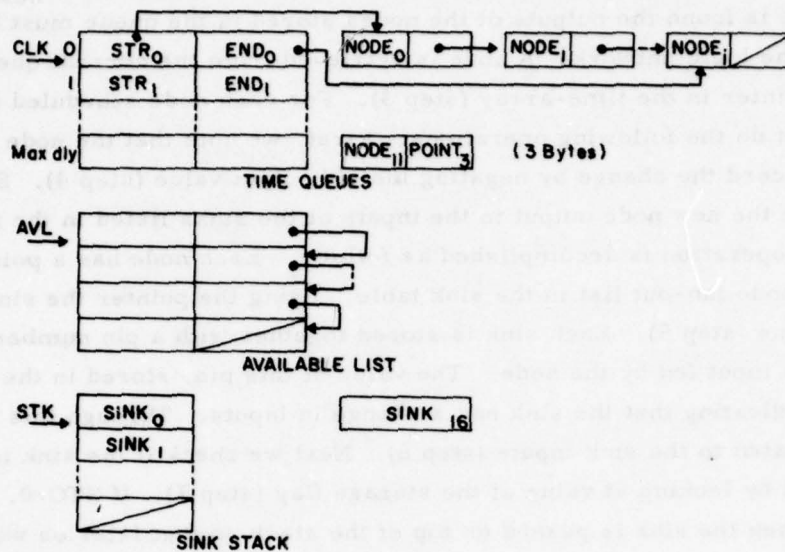


Fig. 2. Time-flow mechanism.

### C. Nominal Delay Algorithm

As mentioned previously, one of the goals of this simulator is to achieve fast response time to input stimuli. We use a table-driven simulator together with selective trace and next-event time-flow mechanism to meet the speed requirements. The simulation algorithm is presented in this section and the major steps are discussed in detail. The steps of the algorithm are listed below.

- (1) if queue pointed by clock is not empty then go to (3).
- (2) clock = clock + 1; go to (1).
- (3) get next node from queue pointed by clock.
- (4) update node output.
- (5) get next node sink from fan-out list.
- (6) propagate node output to sink inputs.
- (7) if sink is stored in stack (STO=1) then go to (9).
- (8) store sink in stack (STC=1).
- (9) if not the last sink go to (5).
- (10) if not the last node in queue go to (3).
- (11) get next sink from stack (STC=0).
- (12) calculate new output of sink.
- (13) if new output  $\neq$  old output then schedule sink in queue pointed by clock + sink delay.
- (14) if stack is not empty then go to (11).
- (15) go to (1).

The algorithm starts by looking for the first non-empty queue (steps 1, 2). When a non-empty queue is found the outputs of the nodes stored in the queue must be propagated throughout the logic network. A node is retrieved from the current queue by using the start of list pointer in the time-array (step 3). For each node scheduled to be processed we must do the following operations. First, we note that the node output has changed and we record the change by negating the old output value (step 4). Second, we must propagate the new node output to the inputs of the sinks listed in the fan-out table. The latter operation is accomplished as follows. Each node has a pointer associated with the node fan-out list in the sink table. Using the pointer the sinks are retrieved one by one (step 5). Each sink is stored together with a pin number corresponding to the sink input fed by the node. The value of this pin, stored in the input table is negated indicating that the sink had a change in inputs. Through this process the node is propagated to the sink inputs (step 6). Next we check if the sink is already stored in the stack by looking at value of the storage flag (step 7). If  $STO=0$ , sink is not in the stack, then the sink is pushed on top of the stack so that later on we can evaluate its output. It is possible that more than one node in the current queue feeds the same sink. We must first determine all the input changes for a sink before we can calculate its output and determine whether and where the sink should be scheduled. Therefore we must gather all the sinks that had changes in inputs due to changes in current nodes' outputs making sure that no sink is stored more than one. To accomplish this operation we make use of a stack (step 8). If we reached the end of the fan-out list for the current node then if the current node is the last one in the current queue then we finished propagating the changes in the nodes' outputs to their sink inputs (steps 9, 10). At this point we release the storage used by the current queue and attach it to the available list for use in the sink scheduling process. For each sink stored in the stack (step 11) we perform the following operations. First, we calculate the new output of the sink. This is done by retrieving the function code and branching to the appropriate output calculation routine. The output of a primitive function such as: AND, OR, NAND, NOR can be calculated with just a few instructions as follows. We examined all the input pins simultaneously and observe that for example the output of an AND gate is 1 if all the inputs are 1's and 0 otherwise, similarly the output of an OR gate is 0 if all the inputs are 0's and 1 otherwise. To calculate the output of a NAND gate we simply negate the inputs and branch to the OR function routine thus savings in code can be achieved (step 12). Next, we compare the old sink output with the new sink output and employ the selective trace technique. If the two outputs are equal then no further action is taken and the next sink in the stack is processed. If the two outputs differ then the sink must be scheduled in time in order to propagate its output. The scheduling process is done as follows. First, we find the time queue in which the sink must be placed by

adding the clock to the sink propagation delay. Addition is done modulo (maximum delay + 1). The sink is inserted at the end of the appropriate queue (end of list pointer) using storage from the available list (step 13). If the stack is empty then we finished scheduling the sinks and we are ready to process the next queue (steps 14, 15). Note that if any sinks had a  $\emptyset$  propagation delay then the clock is not incremented. We finally note that the logic circuit reaches steady state when all the time-queues become empty.

The basic steps of the algorithm were presented; however, the simulator performs other functions such as keeping track of the size of the queues, updating of the hardware input and output parts, and keeping in synchronization with other model units. Also the simulator must initialize the dynamic tables before any modeling can be performed. All these operations can be performed in a straightforward manner and are not considered in the present report.

#### D. Modeling System Architecture

In this section we will consider the hardware requirements for the modeling system. The basic modules, the model unit, and the control units will be described and the operation procedures presented.

Each subunit in the system will be modeled by a microcomputer contained on one card. Total system simulation is achieved by wiring together the microprocessor cards. Figure 3 shows a typical modeling system consisting of an array of microproces-

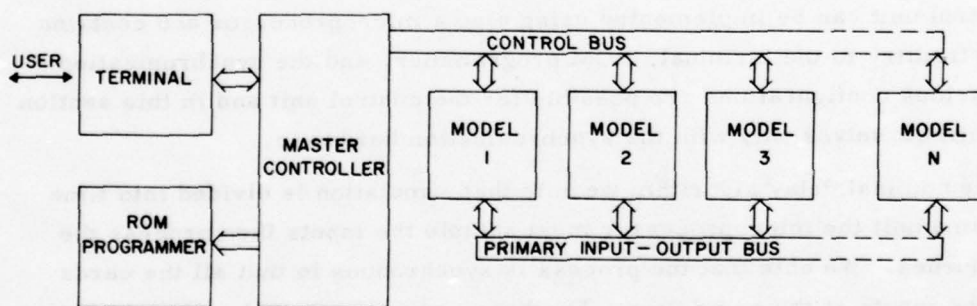


Fig. 3. Modeling system.

sors operating under centralized control. The master controller serves the following functions. It acts as a front-end for the user who interacts with the system. The user feeds in the description of the system to be simulated. The information is transferred by the master controller to a read-only memory programmer. The read-only memory modules containing the static tables for each subunit are then used to personalize each

microprocessor card. The master controller has the task of synchronizing the microprocessor cards. This is achieved through control lines to be described later on in this section. The system clock which controls the simulation is also generated by the master controller.

The model unit contains the microprocessor (CPU), the read-only memory (ROM) modules with the simulation program and the static tables describing the logic network, the random-access memory for storage of the dynamic tables, time queues and program variables, the primary input and output ports, and the control ports. The architecture of the model unit is presented in Figure 4.

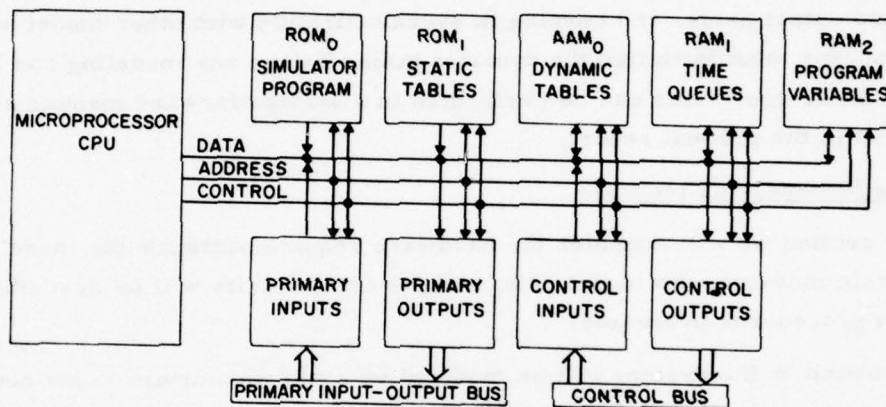


Fig. 4. Model unit.

The control unit can be implemented using also a microprocessor and contains the interface circuitry to the terminal, ROM programmer, and the synchronization hardware. Various configurations are possible for the control unit and in this section we shall concern ourselves only with the synchronization hardware.

Using the nominal delay algorithm we note that simulation is divided into time units. Each time unit the microprocessor must sample the inputs then process the current time queues. We note that the process is synchronous in that all the cards must sample the inputs at the same time. Furthermore we know that a digital system is usually controlled by a system clock. Each clock period consists of several time units during which changes in the logic network are processed. Before a new clock cycle can be started the logic network must reach steady state. With the above considerations in mind we introduce three signal lines sufficient to control the microprocessor array. They are: synchronization line (SYNC), the acknowledge line (ACK), and the steady state line (SS). The SYNC line is generated by the master controller and indicates to each card when to start the next time unit simulation. The ACK line is

generated by each microprocessor and indicates that the simulation of the current time unit is complete. The SS line is generated by each microprocessor and indicates that the subunit simulated has reached steady state. The system clock is generated by the master controller and feeds into the primary inputs of the model unit, the clock is not a control line for the model unit. The synchronization process is shown in Figure 5.

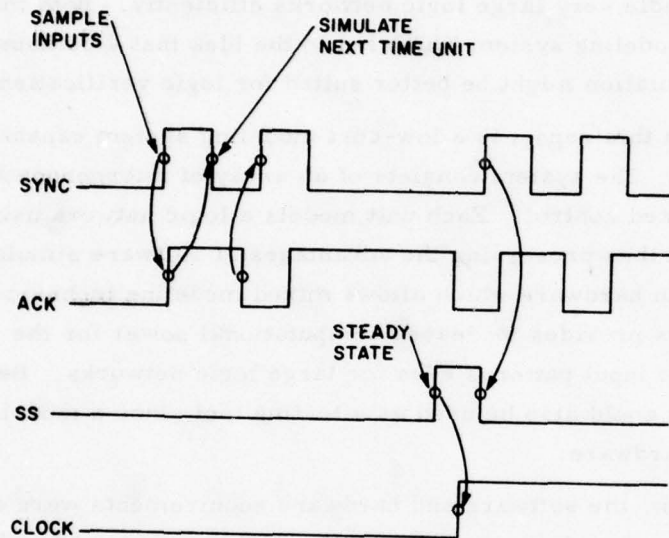


Fig. 5. Synchronization process.

If it is desired to connect the modeling system to "real" hardware then the following limitation should be considered. The "real" hardware should be capable of static operation; there should not be any limit on the speed of operation since the array of microprocessors has a slow response in "real-time." Furthermore the clock of the "real" hardware should be controlled by the master unit to achieve proper synchronization with the array of microprocessors.

So far we have described a basic modeling system for digital systems. The simulation algorithm and the system architecture were presented. The system can be enhanced by the addition of other features which can be implemented in a straightforward manner. For example, it is desirable to have functions to help in the debug process of the logic network. Such functions could be: signal trace capability, set/reset of gate inputs, fault insertion. The new features could be added easily to the simulator structure presented making the modeling system a valuable tool in the design process.

### E. Conclusion

This report presents a new approach towards modeling a large digital system. Due to new advances in semiconductor technology, the size and complexity of digital systems have been increasing at a steady rate. The logic verification phase of the design process has become more complicated due to a lack of adequate tools. It was shown that the two commonly used methods for logic verification, software simulation and breadboarding, cannot handle very large logic networks efficiently. Both methods had advantages desired in a modeling system which led to the idea that a combination of hardware and software simulation might be better suited for logic verification.

The system described in this report is a low-cost modeling system capable of fast response and high interaction. The system consists of an array of microprocessor units operating under centralized control. Each unit models a logic network using a software simulation algorithm thus preserving the advantages of software simulation. The units are interconnected in hardware which allows mixed modeling techniques. The use of multiple processors provides increased computational power for the system which leads to fast response to input patterns even for large logic networks. Besides modeling purposes the system could also be used as a testing tool since a model unit could be replaced by "real" hardware.

To implement the system, the software and hardware requirements were described. The nominal delay algorithm analyzed, used standard simulation techniques such as: function simulation, selective trace and next-event time flow mechanism. The algorithm and simulation data base were presented in a form suitable for implementation on a microprocessor. The basic system architecture consisting of the control unit, model unit, and synchronization hardware was described.

This brief report is not meant to be a complete description of a modeling system but provides a new approach towards logic verification and the basic concepts for implementation. Obviously there is room for improvement in that other features such as debug aids should be added. Additions to and modifications of the system for particular applications should be straightforward to implement on the basic structure presented.

D.R. Kaufman and E.J. Smith

### REFERENCES

1. S. Seshu, "An Improved Diagnosis Program," IEEE Trans. EC-14, pp. 76-79 (January 1965).
2. G.G. Hays, "Computer-Aided Design: Simulation of Digital Design Logic," IEEE Trans. C-18, pp. 1-10 (January 1969).

3. S. A. Szigenda, D. Rouse and E. Thompson, "A Model and Implementation of a Universal Time Delay Simulator for Large Digital Nets," AFIPS Con. Proc., V-36, pp. 207-246 (1970).
4. S. A. Szigenda and E. Thompson, "Digital Logic Simulation in a Time-Based, Table-Driven Environment," Computer, pp. 24-49 (March 1975).
5. S. A. Szigenda and E. Thompson, "Modeling and Digital Simulation for Design Verification and Diagnosis," IEEE Trans. C-25, pp. 1242-1253 (December 1976).
6. H. Y. Chang and S. G. Chappell, "Deductive Techniques for Simulating Logic Circuits," Computer, pp. 52-59 (March 1975).
7. D. E. Kunth, "Fundamental Algorithms," Addison-Wesley (1973).
8. F. J. Hill and G. R. Peterson, "Switching Theory and Logical Design," Wiley (1974).

## ADAPTIVE CHANNEL CAPACITY CONTROLLERS FOR COMMUNICATION NETWORKS

L. Shaw and K. Sohraby

This study considers the performance improvements which might be attained by switching the channel capacity for messages leaving a buffer in a store-and-forward node of a communication network. The proposed controller bases its decision on the number of messages currently in the buffer. Such decisions are made relatively frequently, say after intervals during which a small number of messages are expected to arrive and depart. This mode of operation is feasible where several priority classes of messages share transmission links (e.g., voice and data) and the relative share for each class can be quickly changed.

The analysis makes use of transient properties of M/M/1 queues, approximation of Non-Poisson queues by Poisson queues, Dynamic Programming, special search techniques for minimizing multimodal objective functions, and hierarchical optimization for decoupling the network optimization into a sequence of single queue optimizations.

This project was introduced in Ref. <sup>2</sup>, and is more completely described in a Ph. D. thesis. <sup>1</sup>

A. Single Queue Optimization

A channel capacity  $\mu \in \{\mu_1, \dots, \mu_m\} = U$  is to be chosen for the time interval  $0 < t \leq T$  in order to minimize the following expressions for mean delay per message plus channel utilization cost, when there are  $i$  messages in the queue at  $t = 0$ .

$$C_i = \int_0^T \left[ \sum_{j=1}^{K-1} j p_{ij}(\mu, t) + \beta p_{iK}(\mu, t) \right] dt + \alpha T \rho(\mu) \quad (1)$$

The  $p_{ij}(t)$  are transition probabilities for a queue with Poisson arrivals (rate  $\lambda$ ), exponential service times (rate  $\mu$ ), waiting room of size  $(K-1)$ . These probabilities are obtained by integration of the birth - death differential equations. The cost of lost messages is introduced by the factor  $\beta$ , the channel costs are  $\rho(\mu)$  per second, and  $\alpha$  represents a weighting factor between dollars of channel cost and seconds of message delay.

Since  $U$  is finite there is clearly an optimal policy  $\underline{M}$  which assigns some  $\mu_j$  to each  $M_i$  so as to minimize  $C_i$  ( $i=0, 1, 2, \dots, (K-1)$ ). This best policy can be found efficiently by Howard's policy improvement method of Dynamic Programming.

The output process of such a queue does not have a simple structure. However, it can be well approximated in each  $T$  second interval by a Poisson process. For

each initial state  $i$  and service rate  $\mu$ , one can compute the mean departure rate over the  $T$  second interval as

$$\eta_i(\mu) = \mu \left[ 1 - \frac{1}{T} \int_0^T p_{i0}(t) dt \right] . \quad (2)$$

where  $p_{i0}(t)$  is the transient probability for an  $M/M/1/K$  queue.

The effectiveness of this poissonian assumption was tested by simulation of both the actual and approximating processes as the inputs to a second queue. The average message delays in the second queue served as the relevant measure of approximation quality. The resulting degree of approximation depended on the initial state of the second queue, numerical values of basic arrival and departure rates, etc. An average approximation error of about 10% justified use of this method.

#### B. Multilevel Optimization of a Network of Adaptive Capacity Nodes

The criterion function for a network of adaptive capacity nodes is the sum of delays and channel utilization costs for each node. For a network of  $N$  nodes

$$\bar{C} = \sum_{i=1}^N \bar{C}_i \quad (3)$$

Unfortunately, the capacity assignment policies for the various nodes cannot be optimized separately because policy changes at one node will change the arrival rates at the following nodes. However, the additive form for the network cost suggests consideration of hierarchical optimization techniques which have been effective in similar optimization problems.<sup>3</sup>

The simple three node loop network in Fig. 1 displays the interaction of individual node optimizations. Reference 1 explains how the hierarchical optimization described here for this simple network can be generalized in a straightforward manner to any network

The  $i^{\text{th}}$  node is characterized by

$K_i$  = capacity of the  $i^{\text{th}}$  buffer (messages)

$\underline{M}_i$  = control policy vector with  $j^{\text{th}}$  component

$M_{ij}$  = capacity when  $j$  messages are present at the beginning of the control interval.

$\underline{X}_i$  = input vector with components

$\eta_{ij} = x_{i, 2j} = j^{\text{th}}$  possible arrival rate (Poisson) ( $j = 1, 2 \dots R_i$ )

$\pi_{ij} = x_{i, 2j-1} =$  probability of occurrence of  $j^{\text{th}}$  possible arrival rate

$\underline{Z}_i = T(\underline{X}_i, \underline{M}_i) =$  vector of output rates and their probabilities.

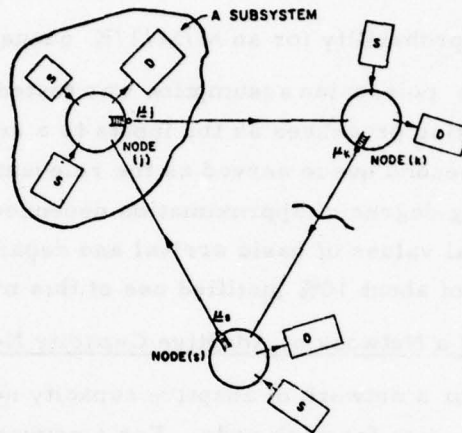


Fig. 1. A typical 3-node network with its subsystems.

The definition of  $\underline{M}_i$  implies that the controller at each node "sees" only his own state. Other possibilities which have been considered<sup>1</sup> include i) a controller which sees the states at its node and at all nodes which feed it; and ii) a centralized controller which uses the states of the queues at all nodes in the network.

We want to choose  $\underline{M}_i$  ( $i = 1, 2 \dots N$ ) to minimize  $\bar{C}$ . The hierarchical approach defines a sequence of subgoals for each node. The second-level computation revises the subgoals, after examining the results of the  $N$  previous subgoal minimizations, in a manner which aims the procedure toward the global network optimization.

The first-level imagines each node to be isolated and free to choose its channel policy  $\underline{M}_i$  and its arrival process. However, connection with the real problem is maintained by augmenting the local cost to

$$\tilde{C}_i = \bar{C}_i + \underline{P}_{i+1}^T \underline{Z}_i - \underline{P}_i^T \underline{X}_i \quad (4)$$

so that the modified total network cost becomes

$$\begin{aligned}\tilde{C} &= \sum_{i=1}^N \tilde{C}_i \\ \tilde{C} &= \bar{C} + \sum_{i=1}^N \underline{P}_i^T (\underline{Z}_{i-1} - \underline{X}_i)\end{aligned}\quad (5)$$

The  $\underline{P}_i$  vectors are Lagrange multipliers or penalty factors which are chosen by the second-level computation to penalize discrepancies from the constraint  $\underline{Z}_{i-1} = \underline{X}_i$ , corresponding to the network connection. (Note that the loop constraint is incorporated by the notational definitions  $\underline{P}_{N+1} = \underline{P}_1$  and  $\underline{Z}_0 = \underline{Z}_N$ .) Clearly, the  $\{\underline{M}_i\}$  corresponding to the global optimum (Min  $\bar{C}$ ) will also make  $\tilde{C} = \bar{C}$ .

The hierarchical optimization procedure is summarized as follows.

1. Given an initial set  $\{\underline{P}_i\}$ , find

$$\begin{aligned}\tilde{C}_i[1] &= \text{Min}_{\underline{X}_i, \underline{M}_i} \tilde{C}_i \text{ corresponding to } \underline{X}_i[1], \underline{M}_i[1] \\ &\text{and resulting in } \underline{Z}_i[1].\end{aligned}$$

2. Modify the  $\underline{P}_i$ , In order to increase the weight on bit input-output discrepancies, we can use

$$\underline{P}_i[2] = \underline{P}_i[1] + \Delta (\underline{Z}_{i-1}[1]) \quad (6)$$

As described in Ref. 3 other second-level penalty adjustments may be more efficient.

3. Return to step #1 and compute  $\underline{X}_i[2]$ ,  $\underline{M}_i[2]$ ,  $\underline{Z}_i[2]$  using  $\{\underline{P}_i[2]\}$ , etc.

This kind of procedure can be shown to converge if the original optimization problem satisfies suitable conditions such as convexity of the cost function.<sup>3</sup> Verification of those conditions can be as complicated as solving the original optimization problem, but experiments with this approach on this network problem have converged. Even if complete convergence does not occur, these steps should produce reasonable policies whose effectiveness can be checked directly.

The network problem requires imposition of physical constraints on the variables in step #1 of the optimization.

1.  $M_{ij} \in \{\mu_{i1}, \mu_{i2}, \dots, \mu_{ik_i}\}$  = set of possible channel capacities.

2.  $0 \leq x_{i,2j-1} < (\eta_{ij})_{\max}$  (7)

3.  $0 \leq x_{i,2j-1} \leq 1$ ,  $\sum_j x_{i,2j-1} = 1$

Moreover, the mean delay part of the cost in Eq. (1) must be modified as follows to account for the random selection of Poisson arrival processes.

Using the notation  $p_{ij}^m(\mu, t)$  for the transition probability from state  $i$  to state  $j$  in  $t$  seconds with arrival rate  $\eta_{im}$  and departure rate  $\mu$ , the generalized form of mean delay plus channel cost at node  $i$  becomes

$$\bar{C}_i(\mu) = \sum_{m=1}^{K_i} \pi_{im} \int_0^T \left[ \sum_{j=1}^{K_i-1} p_{ij}^m(\mu, t) + \beta p_{ik}^m(\mu, t) \right] dt + \alpha T C(\mu) \quad (8)$$

Similarly, the effective departure rate in Eq. (2) must be generalized to

$$\eta_i(\mu) = \mu \left[ 1 - \frac{1}{T} \sum_{m=1}^{K_i-1} \pi_{im} \int_0^T p_{io}^m(\mu, t) dt \right] \quad (9)$$

Minimization of  $\bar{C}_i$  in Eq. (4) using Eq. (8) is carried out in two steps. With  $\underline{X}_i$  fixed, the minimization with respect to channel policy  $\underline{M}_i$  is carried out by the Dynamic Programming approach mentioned earlier. This requires, for each candidate  $\underline{M}_i$ , integration of the transient probabilities in Eq. (8) for each arrival rate  $\eta_{im}$ , as well as in Equation (9). In the long run, the cost at node  $j$  per  $T$  seconds will be a number  $\bar{C}_j$ , independent of the initial state  $i$ . This is the quantity which is augmented by penalty costs in Equation (4). This averaging approach contains the further implicit assumption (clearly an approximation) that the arrival rates in separate  $T$ -second intervals are mutually independent.

The other half of the first level minimization requires a search over all admissible input vectors  $\underline{X}_i$ . The procedure of Chichinadze<sup>4</sup> was used to seek the global minimum of a multivariable objective function which has several local minima. This method is most easily explained via minimization of an  $S(x)$  where  $x$  is a scalar. A function  $\psi(\xi)$  is defined as the total width of  $x$ -intervals in which  $S(x) < \xi$ . A curve passing through  $\psi(\xi)$  for several values  $\xi$  can be extrapolated to find

$$\xi_m = \min_x S(x)$$

where  $\psi(\xi_m) = 0$ . With  $\varphi(\xi_m)$  as the center of gravity of the  $x$ -intervals where  $S(x) < \xi$ ,  $\varphi(\xi)$  can be extrapolated to  $\varphi(\xi_m) = x^*$  to estimate the optimizing value of  $x$ . Both  $\psi(\xi)$  and  $\varphi(\xi)$  are estimated by computing  $S(x)$  for many values of  $x$ .

This first-level optimization is very time consuming, since it incorporates integration of the transient queue behavior and searching for the minimum of a multimodal function of a large number of variables. In the three node loop network example below, where each node stores a maximum of 4 messages, the entire hierarchical optimization (involving five iterations of the  $\{P_i\}$ ) took about 28 minutes on an IBM 360/65.

Example

For an example of the design of the above three node network, the storage capacities of buffers are assumed to be  $K_1 = K_2 = K_3 = 4$  messages (5 states for each buffer), and at each node, an arriving message has a probability of  $1/2$  of departing from the network and of  $1/2$  of queuing for transmission to the assumed next node. The arrivals from outside the network have rates  $\lambda_1 = 1$ ,  $\lambda_2 = 2$ ,  $\lambda_3 = 1.5$  messages per minute, and the control interval is taken to be  $T = 2$  minutes. The delay control parameter for each buffer is assumed  $\alpha = 0.4$  and the blocking parameter  $\beta = 10$ . A linear channel cost of the form  $C(\mu_i) = \mu_i$  is used for all the buffers.

A value of  $\Delta = 0.1$  was used in the penalty updating of Eq. (6), and 5 iterations of that updating were required for effective convergence. The final iteration results of channel capacity policies as well as extrapolated and actual performance costs are listed in the following table:

TABLE I. Adaptive channel results.

Node	Channel Capacity Policy (Kb/sec) States (0, 1, 2, 3, 4)	Actual performance Cost	Extrapolated performance Cost: $\xi_m$
1	2.4, 2.4, 3.6, 6, 7.2	10.18	9.65
2	3.6, 4.8, 6, 7.2, 7.8	10.23	9.36
3	4.8, 7.2, 9.6, 9.6, 10.8	15.19	14.71
		35.60 Total	

As comparison, the design of this three node network with optimal fixed channel capacities and finite storage buffers ( $K_1 = K_2 = K_3 = 4$  messages) was considered. This problem was solved through multilevel optimization techniques of adaptive networks. The following table presents fixed channel capacities as well as performance costs of each node;

TABLE II. Fixed channel results.

Node	Fixed Channel (Kb/sec)	Performance Cost
1	6	12.61
2	10.8	17.96
3	10.8	16.72
		47.29 Total

This example reveals about a 30% total performance cost improvement by the adaptive network model.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

L. Shaw, K. Sohraby

#### REFERENCES

1. K. Sohraby, "Adaptive Channel Capacity, Distributed Message-Switching Communication Networks," Ph.D (EE) Dissertation, Polytechnic Institute of New York (June 1978).
2. K. Sohraby, L. Shaw, R.R. Boorstyn, "Communication Networks with Adaptive Channel Capacities," Progress Report No. 42 to JSTAC, Polytechnic Institute of New York, Report No. R-452.42-77, pp. 399-407 (1977).
3. L.S. Lasdon "Optimization Theory for Large Systems, (New York: The Millan Co., 1970).
4. V.K. Chichinadze, "The  $\psi$  - Transform for Solving Linear and Nonlinear Programming Problems," Automatica, Vol. 5, pp. 347-355 (1969).

TABLE I. Adaptive Channel Capacity

Node	Channel Capacity (KHz/sec)	Cost
1	10.0	1.0
2	10.0	1.0
3	10.0	1.0
4	10.0	1.0
5	10.0	1.0
6	10.0	1.0
7	10.0	1.0
8	10.0	1.0
9	10.0	1.0
10	10.0	1.0
11	10.0	1.0
12	10.0	1.0
13	10.0	1.0
14	10.0	1.0
15	10.0	1.0
16	10.0	1.0
17	10.0	1.0
18	10.0	1.0
19	10.0	1.0
20	10.0	1.0
21	10.0	1.0
22	10.0	1.0
23	10.0	1.0
24	10.0	1.0
25	10.0	1.0
26	10.0	1.0
27	10.0	1.0
28	10.0	1.0
29	10.0	1.0
30	10.0	1.0
31	10.0	1.0
32	10.0	1.0
33	10.0	1.0
34	10.0	1.0
35	10.0	1.0
36	10.0	1.0
37	10.0	1.0
38	10.0	1.0
39	10.0	1.0
40	10.0	1.0
41	10.0	1.0
42	10.0	1.0
43	10.0	1.0
44	10.0	1.0
45	10.0	1.0
46	10.0	1.0
47	10.0	1.0
48	10.0	1.0
49	10.0	1.0
50	10.0	1.0
51	10.0	1.0
52	10.0	1.0
53	10.0	1.0
54	10.0	1.0
55	10.0	1.0
56	10.0	1.0
57	10.0	1.0
58	10.0	1.0
59	10.0	1.0
60	10.0	1.0
61	10.0	1.0
62	10.0	1.0
63	10.0	1.0
64	10.0	1.0
65	10.0	1.0
66	10.0	1.0
67	10.0	1.0
68	10.0	1.0
69	10.0	1.0
70	10.0	1.0
71	10.0	1.0
72	10.0	1.0
73	10.0	1.0
74	10.0	1.0
75	10.0	1.0
76	10.0	1.0
77	10.0	1.0
78	10.0	1.0
79	10.0	1.0
80	10.0	1.0
81	10.0	1.0
82	10.0	1.0
83	10.0	1.0
84	10.0	1.0
85	10.0	1.0
86	10.0	1.0
87	10.0	1.0
88	10.0	1.0
89	10.0	1.0
90	10.0	1.0
91	10.0	1.0
92	10.0	1.0
93	10.0	1.0
94	10.0	1.0
95	10.0	1.0
96	10.0	1.0
97	10.0	1.0
98	10.0	1.0
99	10.0	1.0
100	10.0	1.0
Total	1000.0	100.0

TABLE II. Fixed Channel Capacity

Node	Channel Capacity (KHz/sec)	Cost
1	10.0	1.0
2	10.0	1.0
3	10.0	1.0
4	10.0	1.0
5	10.0	1.0
6	10.0	1.0
7	10.0	1.0
8	10.0	1.0
9	10.0	1.0
10	10.0	1.0
11	10.0	1.0
12	10.0	1.0
13	10.0	1.0
14	10.0	1.0
15	10.0	1.0
16	10.0	1.0
17	10.0	1.0
18	10.0	1.0
19	10.0	1.0
20	10.0	1.0
21	10.0	1.0
22	10.0	1.0
23	10.0	1.0
24	10.0	1.0
25	10.0	1.0
26	10.0	1.0
27	10.0	1.0
28	10.0	1.0
29	10.0	1.0
30	10.0	1.0
31	10.0	1.0
32	10.0	1.0
33	10.0	1.0
34	10.0	1.0
35	10.0	1.0
36	10.0	1.0
37	10.0	1.0
38	10.0	1.0
39	10.0	1.0
40	10.0	1.0
41	10.0	1.0
42	10.0	1.0
43	10.0	1.0
44	10.0	1.0
45	10.0	1.0
46	10.0	1.0
47	10.0	1.0
48	10.0	1.0
49	10.0	1.0
50	10.0	1.0
51	10.0	1.0
52	10.0	1.0
53	10.0	1.0
54	10.0	1.0
55	10.0	1.0
56	10.0	1.0
57	10.0	1.0
58	10.0	1.0
59	10.0	1.0
60	10.0	1.0
61	10.0	1.0
62	10.0	1.0
63	10.0	1.0
64	10.0	1.0
65	10.0	1.0
66	10.0	1.0
67	10.0	1.0
68	10.0	1.0
69	10.0	1.0
70	10.0	1.0
71	10.0	1.0
72	10.0	1.0
73	10.0	1.0
74	10.0	1.0
75	10.0	1.0
76	10.0	1.0
77	10.0	1.0
78	10.0	1.0
79	10.0	1.0
80	10.0	1.0
81	10.0	1.0
82	10.0	1.0
83	10.0	1.0
84	10.0	1.0
85	10.0	1.0
86	10.0	1.0
87	10.0	1.0
88	10.0	1.0
89	10.0	1.0
90	10.0	1.0
91	10.0	1.0
92	10.0	1.0
93	10.0	1.0
94	10.0	1.0
95	10.0	1.0
96	10.0	1.0
97	10.0	1.0
98	10.0	1.0
99	10.0	1.0
100	10.0	1.0
Total	1000.0	100.0

## SOFTWARE MODELING STUDIES

M. L. Shooman and H. Ruston

The material presented below describes, in summary fashion, the research results obtained in the modeling of computer software.

A. Major Results

The major thrust of the work is divisible into three areas: software reliability and availability models, test models, and complexity models.

Building upon previous modeling work by Shooman in the reliability area, new macro and micro reliability models have been developed, namely:

- (1) a Markov availability model;
- (2) estimator formulas and tagging methodology for estimating the initial and subsequent number of errors;
- (3) the macro model was made more realistic through the incorporation of error generation effects;
- (4) a micro model based upon path traversal of the software.

The approach in the area of testing has been to classify and model as well as to provide test techniques. The research has resulted in:

- (1) a new method of test data selection;
- (2) a classification of types of tests;
- (3) the construction of automatic test drivers;
- (4) a statistical model of testing;
- (5) a model of an exhaustive test.

Complexity measures have been sought which relate problem complexity to programming effort and reliability. The major accomplishments were:

- (1) modest advances in the application of the classical theory of recursive function to software problems (the method has practical drawbacks);
- (2) development of software structural measures such as accessibility and testability;
- (3) experimental verification and theoretical proofs that Zipf's law applies to computer languages;
- (4) development of estimator formulas for computing the Zipf's law complexity measure.

B. Applications to Software Engineering1. Design Phase

Design is inherently an iterative process, involving many separate steps, repeated at different stages of the project. During the initial stage the proposed design and its alternatives are evaluated for feasibility and performance. One of the first steps is

to gauge the overall scope and feasibility of the job by assessing its complexity. The common approach is to equate it with similar past systems, on which data exists, and draw comparisons. Such a technique is imprecise and of diminished usefulness in a military environment, where most new tasks eclipse their predecessors in complexity.

Some of the complexity measures developed (cf. Section E.4) are related to fundamental properties of the problem and should allow the extrapolation of past results into the future. To utilize the complexity model during early design, we tabulate the input and output quantities, and estimate what proportion of the available operators in the chosen language will be used. These values are substituted into a formula which predicts the operator-operand (token) length. One then relies on historical data for the ratios of errors/token and man-hours/token. Although these ratios depend on historical data, the scaling procedure based upon token complexity is objective, repeatable, and has been shown (in our small number of retrospective estimates) to be reasonably accurate.

The reliability model discussed in Section C.3 can be used to roughly predict the software failure rate and mean-time between software errors at the design stage. One of the inputs to this model is the instruction length of the program. This can be estimated from the token length and historical data on the number of instructions/token in the chosen language. An additional parameter needed in the model is the error discovery rate constant. This constant relates the number of occurrences of external operational errors to the number of residual internal errors. (Note: an internal error is a potential one which only becomes an actual external error when the particular data which excites it is processed.) Much of the research described in this report involves techniques for measuring these model parameters and testing the ease and accuracy of application of the above-mentioned formulas (cf. C.2, D.2, F.3, F.5, F.11).

## 2. Testing and Debugging Phase

The contributions to this phase contain several techniques for: selecting test data, calculating the number of tests, and automatic testing. In addition, contributions were made to classifying software by the ease of testing and classifying tests by their degree of completeness.

The technique for selecting test data involves the mapping of constraints on the output into the equivalent set of constraints on the input test data. A description of this technique is given in Section D.3.

The knowledge of the number of tests is essential for bounding the test costs and developing a test plan. The method described in Section D.4 consists of a sequence of stages. First, the control structure of the software is modeled by a graph. Second,

matrices representing the graph structure are constructed. Lastly, matrix manipulations are performed to obtain the desired number of tests.

The automatic testing technique developed in this contract yields a test driver. This driver analyzes the source code and modifies it. The modification allows the generation of test data which exercises the software in a specific manner. The modification is such that each IF statement is associated with either a bit string variable of 0 or 1, representing the two outcomes. Similarly, the passage of a DO WHILE loop is controlled. Lastly, various program modules are exercised for the initial values. The details of this method are given in Section D.5.

The classification of software is based upon accessibility of certain modules, ease of testing such modules, and the use of resources in such tests. The details of this classification are given in Section E.3.

### 3. Operational Phase

Many of the models described previously can also be applied to the operational phase. As an example, the release of large software systems to the field is accompanied by a flurry of software errors. The field debugging process can be controlled by the reliability models described in Section C.3.

Once steady-state prevails in the field, the most significant parameter of software performance is its availability. The Markov availability models described in Section C.1 may be used to predict the steady-state availability of the software.

The immediate predecessor of the operational phase is the acceptance test. This test delineates the transition between development testing and operational deployment. Section F.2 describes one type of such test. The philosophy of the acceptance test divides the acceptance into two parts. The first part involves testing of each feature within the nominal range. The second part involves testing of stressful conditions as well as boundaries of the range. The software is accepted only if a prestated high percentage of the tests are successful.

Section F.2 delineates the suggested procedures to be followed if the software is not accepted.

## C. Software Error, Reliability, and Availability Models

### 1. Introduction

We shall present here brief summaries of work documented in technical reports and papers relating to models of software errors and their effect on operational reliability and availability. These models are needed for a number of technical and management functions, namely:

- (a) To estimate the cost and development time related to the excision of the errors initially residing in the program.
- (b) To provide an analysis technique for one of the three decisions, that is, (1) to continue more testing, (2) to terminate testing and to accept the software, and (3) to abort a hopeless effort.
- (c) To provide a technique for cost-reliability trade-off among competitive designs. Such trade-off calculations are needed, for example, to choose between a time consuming technique which produces software with few errors and a fast technique leading to many errors.
- (d) To provide definitions and techniques for operational data collection and measurement. The collected data is needed for the determination of model parameters and the measurement of operational system reliability.
- (e) To provide quantitative measures of goodness which enable researchers to assess the promise of a new technique. This can be done by contrasting the error content and reliability of software constructed with and without such a technique.

The research supported under this contract has advanced the field by providing:

- (a) a Markov model of availability,
- (b) an estimate of initial error content of software based upon a tagging procedure,
- (c) a software reliability model incorporating errors generated during the debugging process,
- (d) a micro reliability model based upon path traversal of the software.

## 2. Reliability Models

The initial thrust of modeling focussed on: (1) the number of software errors, (2) the rate of their removal during the testing phase, and (3) the operational failure rate resulting from residual errors. This work leads to three phases of models, to be described here.

### Macro Models

The model developed in the first phase portrayed just the gross features of interest. It provided expressions for the number of remaining errors as the debugging process unfolded. Certain reasonable assumptions were made to relate the number of residual errors to the operational system reliability. One of the assumptions postulated the constancy of the sum of the removed errors and the remaining errors.

The model of the first phase was named the macro model, because it only considered the gross or exterior behavior of the removal process. Such a model is analogous to modeling the exterior input-output relationships of the system without regard as to the interior structure of the system. The macro model is described in References 13, 14 and 15.

### Macro Model with Error Generation

The macro model was refined in the second phase. The refinement consisted of incorporating the errors generated during the debugging. The generated errors were modeled in several ways. Their rate of removal was modeled as a function of the project's manpower, giving rise to a related investigation of manpower deployment strategies. The complete development of the error generation model and the effect of deployment strategies is described in Reference 7.

### Micro Model

The completion of the second phase, just described above, indicated that further improvement in the fidelity of the model will necessitate the incorporation of factors related to the structure of the software. The model which was investigated is based upon the program flowcharts (or equivalent representations, e.g., pseudo-code), its paths, and the execution and the error frequencies along each path. Such a model is based on the assumption that the software system can be decomposed into a modest number of mostly independent paths (structured programs generally satisfy this assumption). Any gross failure of the decomposibility or independence features invalidate such a model. The micro model is described in Reference 16.

### 3. Techniques for Estimation of Error Counts

A method has been developed for the estimation of errors residing in software at any stage of its construction. This method is based on either seeding or tagging of errors in the program under construction. The initial idea of seeding comes from statistical techniques for estimating the size of wild life population, and was first adapted by H. Mills.<sup>17</sup> Mills experimented with the addition of seeded bugs to those naturally residing in the program. After partial debugging by another programmer (i.e., one who does not know of the seeded bugs), the bugs found are separated into the seeded and the natural categories. By using the total number of seeded bugs, the number of seeded bugs found, and the number of natural bugs found, Mills obtained estimates for the initial number of natural bugs. A major problem in applying this technique is the difficulty in "manufacturing" realistic bugs for seeding.

To avoid this difficulty it has been suggested,<sup>18</sup> that instead of seeding program errors, a tagging technique be used. This technique is implemented by having the program debugged independently by two or more testers. A different person, acting as the analyst, analyzes the lists of bugs found and identifies common bugs found by two or more programmers. The number of common bugs found by multiple programmers serves the same purpose as the seeded bugs in the former techniques. The disadvantage of the method is the need for additional personnel (i.e., the independent debuggers).

However, this cost is largely offset by the extra bugs that the additional personnel finds.

The tagging method has a number of advantages over the seeding one:

- (a) The difficulty of manufacturing realistic seeded bugs is eliminated.
- (b) None of the testing and debugging funds are used for the finding of known seeded bugs.

Because of the promise of the method, a major effort was undertaken resulting in a comprehensive report.<sup>8</sup> This report derives several estimator formulas for mean and variance for different sample sizes. The report also describes how the method is to be used for the case of two, and for the case of several, debuggers. A small scale experiment has been performed, and the data is presently being analyzed.

The method is especially useful in estimating the initial number of bugs, to provide measures (1) of the cost of testing and (2) of a parameter in the reliability model. It is also useful for obtaining information for management decisions at certain key points of software development.

#### 4. Markov Availability Models

Models have been developed<sup>6</sup> for the determination of availability of an operational (i.e., field deployed) computer software. For many systems the reliability is not the crucial measure, but rather the ratio of up-time to the total operating time, that is, the system availability.

The models developed are based upon Markov processes. The initial upstate assumes the system to be running with undiscovered bugs. When the first software error is discovered, the system enters the second up-state which represents  $n-1$  bugs. The model thus consists of a sequence of up and down states. The transition probabilities between up and down states are essentially the respective software rates. The transition probabilities between up and down states are the probabilities of repair taking place in a given time interval.

The model leads to difference and differential equations for the state probability (e.g., the probability that the system is in a particular up or down state). The reports show solutions for several different cases involving various assumptions on the failure and repair rates.

#### D. Test Models and Techniques

##### 1. Introduction

We shall present here brief summaries of work documented in technical reports and relating to test models and techniques. These models are needed for a number of technical and management functions, namely:

- (a) To obtain parameter values for certain models described in Section 4.
- (b) To classify and model the testing process.
- (c) To suggest new means and strategies of testing.
- (d) To estimate the costs of software testing.
- (e) To manage and measure progress during the test phase of software development.

The research supported has advanced the field by providing:

- (a) A New method of test data selection,
- (b) A classification of types of tests,
- (c) The construction of automatic test drivers,
- (d) A statistical model of testing,
- (e) A model of an exhaustive test.

## 2. Small Scale Testing

In order to explore the verification of theoretical models (e.g., the tagging estimates), and to estimate parameters needed for other models, four programs have been written by student programmers and careful records were kept on their debugging experience. The initial effort and the four programs are described in Ref. 5 (pp. 22-32), along with the reporting scheme.

The four programs were constructed in increasing order of complexity: (1) salary payroll adjustment, (2) roots of a cubic equation, (3) library search, and (4) ballot counting procedure. Each debugger recorded: (1) the number of bugs and their types discovered during each test run, (2) the time required for analysis of each error and the computer time of the run needed to correct the error, and (3) the chronology of removal of inherent bugs and bugs generated during the testing phase.

As the data reduction continues, we intend to utilize the new measure of complexity (i.e., Zipf Law, see Ref. 9) to explore the relationship between complexity and the initial number of program bugs. We also intend to correlate the Zipf's Law measure of complexity with intuitive subjective estimates of relative complexity by programmers (e.g., consensus estimation, see Reference 19).

## 3. Data Selection Studies

The crucial problem in testing is the selection and generation of an appropriate set of test data. This difficult problem has been studied from several viewpoints. One such view will be described here; several other viewpoints appear in the sections to follow.

The method used in this study uses constraints imposed on the variables. Tracing

of all constraints resulted in the determination of bounds on the variables, and thus construction of test input data. The method requires the construction of dependency tables and has been described in Ref. 1 (pp. 10-12) and in Reference 2 (pp. 33-50). A more complete treatment is given in Reference 20. At present the method is very tedious because of a laborious hand-tracing through the dependency tables. Future efforts will attempt to mechanize the effort (e.g., by perhaps writing a program replacing the hand-tracing) and to apply it to modest size programs. Even though the present method is traceable to programs of any size, the tediousness of the hand-tracing makes it impractical for other than small programs.

#### 4. Determination of the Number of Tests

One of the pertinent questions of interest is the number of tests required of software, so as to verify such software to a desired degree of completeness. A complete verification requires perhaps an exhaustive test, that is, a test for all possible sets of inputs. Obviously, such a test is almost always impractically large. Consequently, practical considerations demand lesser levels of testing. Hierarchies of such tests have been categorized (Ref. 21, Chapter 4) and assigned level numbers<sup>\*</sup> from 0 to 4. Type 0 test exercises each program statement. The highest level is an exhaustive test. The in-between levels represent tests more complete than the 0 type but not exhaustive (e.g., level 2 is execution of all paths in the program). Also intermediate levels can be introduced, such as type 1.5 for example, which identify all the levels of the one level and some of the second level.

We have investigated analytical methods for computing the number for type 2 tests. This work was reported in Ref. 2 (pp. 3-11), Ref. 3 (pp. 40-49), and Reference 4 (pp. 10, 14-16). A more detailed discussion is given in a technical report presently in preparation. The method starts with a graph of the control flow of the program. Theorems from graph theory have been applied to determine the number of type 2 tests. This entails visual search through graphs, which is very cumbersome for large problems. The alternate approach is to represent the control graph by a connectivity matrix. Such a matrix can then be manipulated via matrix transformations to yield the desired number. The matrix method is amenable to computer processing.

#### 5. Automatic Test Drivers

Another approach to the generation of test inputs evolved in the construction of a so-called test driver. Such a driver is a generalized program. The input to the driver

---

<sup>\*</sup>Two different scales for levels have been introduced. The earliest scale went from 0 to 3. The newer scale extends from 0 to 4.

is the program to be tested. The driver generates signals necessary for forcing the traversal of each program path.

The initial driver was implemented in LISP and is described in Reference 5 (pp. 14-17). Further work is reported in References 12 and 22. These references describe the progress achieved in the development of the driver. In particular, the present implementation is in PL/I, which allows driver testing of wider classes of programs. The initial version, for example, was restricted to the loopless program; the present version no longer suffers this constraint.

The essence of the algorithm is to associate a binary digit with the predicate (i.e., condition) of each decider (implemented in PL/I by the IF statement). The state of all the deciders can then be represented by a binary number made up of an ordered concatenation of the individual digits, with each digit representing a single decider. The algorithm generates the entire valid set of such binary numbers, and derives the appropriate driving signals from them.

Similar techniques are used to reduce the testing of a DO loop to just the first and the last passes. Such a reduction greatly reduces the run time of a complete test. For the details, the reader is again referred to the cited references.

It should be observed that the assumptions on the control structure being composed of only simple sequences, repetitions, and selections, are satisfied by a structured program. With the present emphasis on usage of structured programming techniques, these assumptions are not very limiting.

#### 6. Statistical Test Models

A statistical test model has been developed which relates different program errors to the input data set (or sets) which excite and thus display a particular error. The model also gives the probability that these errors will cause the program to fail. This work was reported in References 3 (pp. 19-30) and 4 (Section 3.7). A more detailed description of the model appears in Reference 23.

The model assumes that there are  $N$  total possible test inputs (i.e., input data sets). It is further assumed that each input is equally likely to occur either as the tester's or the user's input. A certain number of these inputs, say  $w$ , will result in operational errors. During the development of the program only the subset  $t$  of the  $N$  possible inputs is used to test the program. The model relates  $P_e$ , the probability of the error occurring during the program use, to the parameters  $W$ ,  $N$ , and  $t$ . The model also yields a rectangular grid representing the test process. Each grid point  $(j, i)$  is interpreted as the excitation of the bug  $i$  by the input  $j$ . The graph also portrays the

two situations when (1) a single input excites several bugs, and (2) different inputs excite the same bug. These two situations are displayed by either a horizontal or vertical interconnection of grid points.

A related study utilizes a similar model to evaluate the probability that a particular sequence excites a bug. This work is reported in Reference 4 (Section 3.7).

#### 7. The Exhaustive Test Model

It is obvious that an exhaustive test is prohibitive in time and cost for almost all practical programs. Nevertheless, the number of tests for an exhaustive test plan still represents a useful upper bound on the actual number of tests that will be used in practical testing. These upper bounds have use in comparing the test efforts needed to test different programs.

The knowledge of the number of tests needed for an exhaustive test is needed if one wishes to use test coverage (that is, the number of the tests performed to the number of possible tests), as a figure of merit. Because of these applications, the modeling of an exhaustive test was studied and the results are reported in References 12, 21, 22 and 24.

The exact computation of the number of tests was made in Ref. 24, for a particular example of a small program. In this example, the program extracted the roots of a quadratic equation, and was designed in PDP-8 assembly language. The number of exhaustive tests depended upon the range of possible values for the three coefficients of the quadratic and the 12 bit word length of the computer.

### E. Complexity Models

#### 1. Introduction

The fundamental problem in the development of software is to construct a product of minimum cost which still meets the quality and reliability design specifications. When the product is completed records often show that the development costs greatly exceed the initial projection. A major cause of the poor estimation is the strong influence of complexity and the difficulty in measuring and assessing its effects.

Complexity, to be more specific, has a major effect on:

- (a) development costs,
- (b) development time,
- (c) amount of testing,
- (d) number of bugs,
- (e) memory needs,
- (f) size of host computer,

- (g) skill and number of programmers,
- (h) maintenance,
- (i) life-cycle costs,

and many others. Complexity is a very elusive type of measure, with its importance perceived by everyone but in an ambiguous and fuzzy fashion.

Complexity affects all the above listed developmental parameters in a nonlinear, hard-to-determine way. Clearly one needs a more sophisticated measure than a mere count of the number of lines of code. There are many examples of simple large programs and complex small programs. The problem of complexity is a difficult one as is evidenced by the number of highly competent researchers working on this problem and the relatively few important results published so far.

The research supported under this contract has advanced the field by providing:

- (a) A comprehensive study of classical methods of complexity with emphasis on recursive function theory. The results were theoretically appealing; however, practically difficult to apply and extend.
- (b) An approach in which software complexity is decomposed into simpler software components. One such component is accessibility, which is a measure of the software structure. Another component is testability, which is the ease with which a module may be accessed and tested. A third component is testedness, which measures the frequency of execution during the testing phase. This approach shows promise but more work is needed.
- (c) A measure of complexity based on Zipf's Law and other results in natural language theory. This approach links known results from natural language and information theories to software complexity questions. (The paper on this subject (Ref. 25) has received the best paper award at the Fall 1977 IEEE COMPCON conference.) The results give answers to such questions as the expected length of program given the initial estimate of the number of variables and the types of needed algorithms. The formulas derived from fundamental principles give answers which correlate well with those obtained independently through software science formulas (Reference 27).

## 2. Application of Recursive Function Theory to Program Complexity

The initial study of software complexity focused on the theories of computational complexity, with the intent of extending such theories to realistic programming problems. Recursive function theory was selected as the most promising candidate for such a study.

Strictly speaking, recursive function theory applies solely to mathematical functions. However, the programming solution to a problem often can be viewed as a function or functional. Thus, the first question of interest is how to reduce a general program to a function. In theory, each output variable can be related to input variables, but the practical problem of obtaining the mathematical functional relationship is not an obvious one. This is further complicated in that classical recursive function theory deals with integer variables.

In our work we extended recursive functions to the domain of non-integers and character strings.<sup>10</sup> Unfortunately, the method is difficult to apply and suffers from several limitations. Consequently, we conclude that the method is of limited use for classifying practical problems in terms of their complexity.

### 3. Component Measures for Evaluation of Software Complexity

One approach to measuring complexity is to decompose it into simpler, easier to comprehend, components. Such an approach has been undertaken and in fact, some of the more promising measures can also be applied to determining the quality of software.

The first component developed and studied was named accessibility. Accessibility measures how easily a module can be reached and thus describes the difficulty in testing such a module (Reference 3, pp. 53-55). The second component was named testability. This component used both accessibility as well as a measure of resources required to test a program module (Reference 3, pp. 55-56). The third component was named testedness and was defined as a function of testability and the frequency of execution during the testing phase (Reference 3, pp. 56-57).

Additional work on these concepts is described in References 20 and 26. This work is very promising but more effort is needed on obtaining more components and correlating them with the overall software complexity.

### 4. Language Theory Measures of Complexity

The use of number of statements as a measure of complexity is often used. It is recognized that such a simplistic measure does not truly gauge the program. Often there are programs with a few lines considerably more complex than programs with many lines. As a better measure Halstead<sup>27</sup> suggested operator and operand count. The analogy between operators, operands and verb nouns suggested the application of Zipf's Law from natural languages to programming languages.

The relationships between number of operators and operand types, as well as their frequency of occurrence, was shown to follow the basic Zipf's Law. An extended form of Zipf's Law has been derived which more closely models some of the experimental data.

The theory culminates in an equation which relates operator and operand length to the number of types.<sup>9, 25</sup> If we estimate early in the design the number of input and output variables and operators (which will be used in the program) we can obtain an estimate of the program length.

These results correlate well with the results which Halstead has obtained using software science techniques (Ref. 21, Chapter 3 and Reference 27).

## F. Other Research in Progress

### 1. Introduction

As is usual in any research endeavor, there are a number of tasks which are in different stages of completion. Some of these represent new work, not reported in previous progress summaries. Other work has been reported earlier, but has not as yet reached the point where a comprehensive technical report is justified.

The brief sections given below describe such tasks. These descriptions have been included, to record the ideas and their stage of progress, in order to provide continuity in the follow-up research.

The eleven technical sections of this chapter can be broadly classified into the following categories:

1. Models: Sections F. 2, F. 3, F. 4
2. Complexity: F. 4, F. 5
3. Software Management Tools: F. 6, F. 7
4. Data Collection: F. 8, F. 9, F. 10, F. 11
5. Design Techniques: F. 12

Each of these sections is self-contained with appropriate literature references to related work. It is anticipated that several of these will be developed into comprehensive technical reports, while others will be incorporated in future research.

### 2. Acceptance Test Models

Most of our research on testing has involved the development phase of software production, 11, 12, 22, and 24. This section describes some preliminary work which we have done on the structure of an acceptance test. Such a test is one of the most important milestones in the management of software development. An acceptance test is basically a vehicle whereby the contractor convinces the contractee that the software is good, correct, and should be accepted. The test which we have proposed is of two parts.

1. A group of  $k$  test cases are selected which test each feature or mode of operation of the program. All parameters are selected well within the normal range of operation.
2. A collection of  $n$  test cases are devised such that they include all known extreme and difficult cases which constitute  $n_1$  tests; the remaining  $(n-n_1)$  cases are distributed over the normal range of operation.

The specific test cases are devised by the contractee (or his representative if third party testing is used) and are unknown to the contractor. If all  $k$  cases run successfully, and at least  $x\%$  of the  $n$  test cases, then the software is accepted; however, the contractor is given a short amount of time to fix the  $n(1-x/100)$  errors. If either part 1 fails or part 2 does not meet the required threshold of  $x\%$  successful tests the contractor is given time to improve the software. After the software improvement the acceptance test (i.e., both parts) is rerun with different test cases.

Work is presently continuing on these ideas with regard to identifying the contractor and contractee risks and relating the results to the software reliability.

### 3. Comparison and Test of Software Reliability Models

Major questions in the research and application of software reliability models concerns the basic assumption, range of applicability, and ease of use of one model versus another. A good summary of the models proposed in the past and their comparison is given in References 28 and 29.

Answers to the above questions can be developed in several ways. One obvious approach is the theoretical investigation of the models, their underlying assumptions, and their limitations using mathematical statistics. This approach, taken by several workers (cf. Ref. 14), suffers from drawbacks. In some cases the computations are intractable. In other cases it is difficult to assess the practical impact of the non-existence of an entity (e.g., a particular moment) or the bias of an estimator.

The best overall test of such models is to use them for the analysis of sufficient actual field data with well-known outcome. If the predictions by-and-large agree with the field experience, we will be less concerned with the mathematical quirks. However, we never have field data in sufficient varieties to relieve our concern.

In order to supplement the field data, we propose to generate additional data by simulation.<sup>30</sup> We may fabricate data of particular types to pin-point and study certain model weaknesses. Suppose, for example, that a reliability model is suspected of being insensitive to changes in failure rate in time. We can generate three sets of simulated data: (1) with constant failure, (2) with linearly increasing failure, (3) with exponentially decreasing failure. We can now study how the model responds to these three pure data patterns. Such study is not possible with real data which contains many patterns.

### 4. Application of Graph Theory to Software Reliability

The application of graph theory to problems in software reliability was studied. The motivation for this study is the fact that linear graphs (i.e., vertices connected

by edges in various configurations), suggest themselves as logical structures for representing computer programs. However, the detailed application of these concepts is not obvious. To be sure, some elementary approaches suggest themselves at the outset, and the flowchart concept, so basic to computer program documentation, is very close to a linear graph. For example, the vertices of the graph may be operational symbols of the chart, and similarly, the graph edges are the flow lines of the chart. Alternatively, the vertices may represent states reached in the program, and the edges indicate the passage between stages (see Ref. 21, pp. 3-73ff). Such models can be used to study the path structures in the code. This in turn can relate to the kind of testing required to achieve various levels of software reliability. Covering a graph with paths and related concepts form a standard part of graph theory. The interplay between this material and software testing algorithms has received attention from a number of investigators.

Graph theory may contribute to other aspects of software reliability, as well as forming the basic representational structure just described. Several such aspects have been considered. These are briefly described below for possible future use even though they were not carried beyond the initial stages. These aspects deal with complexity and fundamental characterization of software.

As in the case of hardware, complexity as a concept carries both positive and negative aspects for reliability. On the one hand complex systems may be expected to be less reliable than simple ones; on the other, a major technique for improving reliability is enhancing system design which may increase complexity. It becomes clear that a conceptual use of "complexity" is not satisfactory, but specific and quantifiable definitions are required. In graph theory there are several definitions of graph complexity, the most well-known being cyclomatic complexity (discussed in Ref. 21, pp. 3-73ff and Reference 32). Two other measure of complexity are well defined. One uses the information content in a graph and its underlying automorphic group structure and was developed largely by Monshowitz.<sup>33</sup> Another approach to complexity has been formulated by Minoli and is described in Reference 34. He specifies a set of desirable properties for a complexity measure and defines its mathematical form. These measures of complexity were not developed with software applications in mind. It is likely that more appropriate measures can be produced by starting from basic needs of software analysis.

As another approach to complexity, consider activity on the linear graph or network (representing computer software) rather than the static (logical) structure alone, as is the case for the complexity measures mentioned above. One such approach is due to Flynn and is discussed in detail in his dissertation.<sup>35</sup>

Though most people see at once a connection between software code, flow charts, and linear graphs, deep consideration exposes a lack of fundamental characterization. One wishes to employ mathematics to the study of a variety of computer software problems, but in fact, it is not clear what kind of mathematical object represents a particular computer software. Indeed, the representation may differ depending upon need and intended usage. For example, consider the characterization of complexity. How should one represent the software and then impose a goal oriented definition of complexity upon the resulting mathematical object? In any case, regardless of specific representations, one should have a mathematical characterization of the software. Study indicates that such characterization would be likely to follow set theoretic formulations. This in turn leads to linear graphs which have a direct correspondence to subset structures. This basic characterization, however, has not been completed as yet.

An aid to the study of the interplay between graphs and code is the automatic generation of linear graphs which represent computer code. A program to take FORTRAN code and produce a graph from it was started using PL/I. The procedure looks promising and will be further studied in the future.

#### 5. Correlation of Program Errors with Complexity Measures

In order to assess the validity of complexity measures advanced in Refs. 9 and 27, a data collection has been planned using an IBM 370-125 computer operating under DOS. All output generated by the computer is stored on the output queue of the operating system generated file POWERQ. The process is cumulative and continues until POWERQ is filled, which takes approximately one week.

As a next step the POWERQ output file will be run through a program which scans it for FORTRAN programs. When a FORTRAN program is found, the program copies the job card information, then continues scanning for syntax and run time errors. These are identified by the prefix ILF in the output queue. There are a total of 301 such errors detectable by the system. The errors are copied after the job card information. The process continues until the entire POWERQ file has been scanned. The sought-after information is copied onto a tape.

The resulting tape will then be sorted according to job number, name and program number, which will provide unique identification of the program. After sorting, the tape will be merged with a master tape on which the following information will be accumulated:

1. A count of all runs which contain syntax errors. Syntax errors are those with error numbers from 1 to 206.

2. A count of all runs which contain run time errors. These are identified by error numbers 207 to 301.
3. A count of all error free runs. It will be assumed that the number of logical errors is one less than the number of error free runs.
4. A total count of all runs.

The final version of each program will be collected from the students and read onto a tape using the job card with which it was run. These programs will then be run through the modified version of a program obtained from Professor Halstead, which counts frequency and other parameters.

Using the frequency counts, the entropy measure (i.e.,  $f_i \log_2 p_i$ ) will be obtained and stored, together with job card information, on a tape. The probabilities  $p_i$  will be obtained from:

1. previously run programs
2. the current programs which are being tested.

The tape so generated will eventually be sorted. This tape, containing the complexity measure, will then be compared to the error counts, and a correlation determined, if one exists.

Most of the programming effort to date has been spent on the program needed to handle the POWERQ file. Methods were studied to access and process this file. The program is now almost ready and only minor changes remain. Further work will be described in the future.

#### 6. Choice of Strategy in Software Revision

For a wide variety of reasons, software undergoes changes and revisions during its life cycle. Examples of such reasons are:

1. A new or changed specification has to be incorporated.
2. The host computer or source language for the software has been changed.
3. Peripheral equipment has been changed.
4. Significant bugs are found in deployed software.
5. Format changes in input or output are required.
6. The program is to be used as a module in different software and the interfaces must be matched.

When confronted with such changes, the first decision concerns the choice among the following approaches:

1. should the code be discarded and a new code written
2. should the code be retained, but major modification be made,

3. should only minor modifications be made.

The choice among these three alternatives is predicated upon: the quality of the software under consideration, the extent and significance of the required change, the expected life-time of the new software, and the budgeted resources. The selection of alternatives has to be made on a technical and managerial basis. At present there are few objective tools to aid in such decisions.

Early work in this area by S. Amster<sup>36</sup> correlated various subjective and objective measures of program quality. The subjective measures elicited opinions on the extent of changes needed to improve clarity, decrease memory size, and decrease run-time. The objective measures included such factors as number of lines of code, number of GoTo's, number of calls, and so on.

More recently M. Halstead (Ref. 27, Chapter 7) has classified several so-called program "impurities," which detract from clarity. The elimination of these impurities improves the program readability.

We are investigating the use of Amster's and Halstead's techniques, as well as others, to produce a ranking to aid the designer and the manager in their decision.

#### 7. Effect of Organizational Structure on Software Development

On all large technological tasks, the organizational structure of the project team has a large influence on productivity, reliability, and the quality of the product. These effects are especially strongly exhibited in software development.<sup>39</sup> A major phenomenon of interest is that productivity is often not proportional to man hours. This fact has been observed by many and articulated by Brooks.<sup>37</sup> In fact, at some stages of software development, adding workers to the project slows the pace of progress.

We have initiated research in this area by studying the research literature on graph theoretic models of organizational behavior. Much of the qualitative literature relates to research done by psychologists prior to 1950. This work, which supplies graph structures for various group organizational structures, is summarized in Reference 38.

A simplistic assumption is usually made that programming productivity is a direct function of charged time. However, this is not a satisfactory assumption because charged time represents raw man hours composed of personal time (coffee breaks, conversations, etc.), studying time, communication time, and lastly, productive time. We can assume that the proportion of personal time is fixed at say 10% regardless of the organizational structure. However, the remaining time divisions are highly dependent on the organizational structure. Preliminary results indicate<sup>40</sup> that we are able

to model the communication links as paths in the graph structure, and the constraints (a fixed number of man hours or a fixed number of programmers). The object is to evaluate the merits of various organization schemes so as to minimize the need for communication and thus increase the productivity.

A simple example illustrates the phenomenon of increased man hours resulting in decreased productivity. Suppose 10 men work on a project, and their total of 400 hours per week are spent on 40 hours of personal time, 160 hours of studying and communications time, and 200 hours of productive time. A programmer is added to the group, and since he is new, he spends his 40 hours on 4 hours of personal time, 16 hours of productive time, and 20 hours on studying and communications. Thus, he adds 16 hours of productive time to the task. However, suppose that the integration of this new man into the group requires 1 hour review of the project attended by all. This is a group loss of 10 hours of productive time. In addition, the new man is assigned to spend a day (8 hours) with an experienced team member to help him get started on his assigned task. Thus, another 8 hours are lost, and the net result of this added worker is a two hour loss.

Work is presently continuing on establishing further relationships among the organizational structure and productivity, reliability and quality, and of designing a realistic model of this task.

#### 8. Collection, Storage, and Retrieval of Software Reliability Data

The collection, storage, and availability of reliability data is of vital importance to the worker in the field of software reliability. The data is needed by experimentalists to suggest models. It is needed by theoreticians to test the hypothesis of their postulated models and to evaluate their parameters. Finally, it is needed by managers and designers as a data base for design decisions.

A preliminary study supported by RADC<sup>31</sup> discusses the scope, specific techniques, and feasibility of establishing a software reliability data base. In the course of writing material on software engineering,<sup>21</sup> the following rough estimates were made on the amount of information generated annually in writing software for the Air Force.

- a. The total number of bugs/year is roughly between  $7.5 * 10^5$  and  $7.5 * 10^6$ .
- b. Assuming that ten data descriptors/bug must be stored, one needs between  $7.5 * 10^6$  and  $7.5 * 10^7$  words of storage. All these words may be stored on a few reels of tape and disks.
- c. Assuming microfiche is used for permanent backup storage, one needs between  $7.5 * 10^3$  and  $7.5 * 10^4$  microfiche.

Further details and the underlying assumptions are given in Reference 21.

### 9. Classification and Enumeration of Program Errors

In order to obtain some preliminary quantitative data on the various types of programming errors, approximately 1,000 programs were analyzed manually. In this work, described in Ref. 1, we classified the errors into three broad categories, namely, syntax, semantic, and algorithmic. Within these three categories, we enumerated 117 different errors. Rather than listing individual numbers of occurrences, we rated the errors on the scale from 1 (typifying a rare occurrence) to 5 (for a frequent occurrence).

The desired statistics were obtained from programs in a first and in a second programming course for the following reasons.

1. The student programs were under our control.
2. In a typical programming assignment given to a class of 25 students, the same program is run approximately 50-100 times (about 2-4 runs per student, with the extra runs for debugging or cosmetic improvements). Since the assignments are known, we know the results and can note the presence of algorithmic errors, a task that will be impossible in a "strange" program.
3. The student programs are a good sample because the number of errors is nearly constant in a programming assignment. This is so because at the start of the semester, even though the students are inexperienced, the assignments are simple and have a few lines of code. Later in the semester the students are more experienced and make fewer errors per line of code, but the assignments are harder and require more lines of code.

We tried to gather various statistics on programming errors by issuing forms to students and asking them to record various program bugs. However, this method was not successful. Some students were suspicious that these forms were used for grading purposes. Other students did not want to admit to a high number of errors or runs. Consequently, we abandoned this voluntary method. What we did instead was to modify the output for the student programs so as to produce two identical printouts of each program. One was returned to the student, and the other was retained for us in the computer center. We collected these duplicate programs and used them in our analysis. This duplication was made possible by the fact that the students' programs, written in the PL/I language, are compiled with the PLAGO interpreter (which implements a subset of PL/I) developed here and completely under our control.

To compile the statistics we used the form described in Reference 1. Such forms were compiled for each student. From these forms we enumerated the number of different errors. For details we again refer to Reference 1.

The tediousness of the manual analysis and the large volume of programs to be recorded lead to plans for automatic collection. Two such plans are described in the next

two sections.

#### 10. Modifications of the PLAGO Interpreter

The manual classification of compile and execution time errors was discussed in the previous section. This work indicated the need for automatic collection for the following two main reasons:

1. The volume of errors was too large for manual collection and reduction.
2. The students' reporting data was not always reliable. These forms often reported fewer errors than actually occurred, because they thought that such good results would please the instructor.

Consequently an automatic collection scheme was undertaken. The purpose of the collection was to obtain the following information:

1. The number of PLAGO programs submitted. PLAGO is a PL/I dialect containing a scientific subset of PL/I, developed at the Polytechnic and used in our programming courses. PLAGO was chosen as the subject language, because it is entirely in our control.
2. The number of errors in each category. Such recording is possible, because each category is identified by an error number. These categories include both compile and execution time errors.
3. Certain logical errors occurring during execution. This can be done by requiring students to use prescribed variable names in programming assignments and recording the results which are different from the correct ones.

All these tasks were undertaken as a Master Thesis by a student. Unfortunately, he took a leave of absence which continues at this time.

A similar effort, using the FORTRAN compiler, is described in Section F.5. Nevertheless, we hope to continue PLAGO error collection in the future either with the returning student or with other personnel.

#### 11. Small Scale Tests

In order to study, verify, and measure parameters of the theoretical models developed,<sup>8, 9, 14</sup> small scale tests were planned and begun. The tests consisted of four tasks, programmed by student programmers, with careful records kept on their debugging experience. The four tasks and the reporting form are described in Reference 5 (pp. 22-31).

The student programmers were of sophomore-junior standing, with high interest and ability in programming topics. Because of this selectivity we believe their product to be equivalent to programs produced by programmers with intermediate experience. Consequently, we consider the obtained test data to be representative of normal practice.

The programmers were made aware of the importance of maintaining careful and truthful records. They were also given the following specific instructions:

1. The problems were to be analyzed and coded, with both analysis and coding times recorded.
2. The initial program was to be corrected of just the syntax errors. Their number, number of runs needed for their correction, and run times were to be recorded. All printouts were to be saved and numbered.
3. The programs were presented to us (i.e., to either M. Shooman or H. Ruston) for inspection. We then asked the program author and other members of the group to debug each copy independently, recording:
  - a. Number of bugs and types found in each debugging shot.
  - b. Analysis time and computer time for each debugging shot.
  - c. History of removed bugs and generated bugs.
4. A programmer who reached a blind alley had to consult with us. He could neither ask for other help, nor abort the program.
5. The programs were constructed with the following constraints:
  - a. Structured design was to be followed.
  - b. The instructions were to contain no impurities (Reference 27).
  - c. The main program was to be the control structure, with calls to modules (i.e., blocks or procedures) for various tasks.
  - d. No module was to exceed 50 lines.

The first three programs were written by five different student programmers. Most have been completely debugged and documented. We must still debug the remaining programs, record data on the fourth program, and analyze the data.

## 12. Automatic Programming Techniques

This study addresses itself to three questions, namely,

- a. Can a better applications program design be obtained if certain program blocks are available in a computer library? (By better we mean all the goodness factors of a design - less cost, more reliability, etc., however, we do not necessarily count run time or memory size as long as they are within reasons).
  - b. If the answer to the first question is in the affirmative, what blocks should be written and placed in the library,
- and,
- c. how is the information about these blocks to be transmitted to the program designer?

In the research to be described, we assume that the answers to these three questions are:

- a. If program blocks are already available, the incorporation of these blocks certainly simplifies the design task. This statement is evidenced by the fact that to design around available code is the standard design technique.

A further advantage of reusing available code is that inexperienced junior programmers may produce reliable software, if critical parts have been designed by better or more experienced programmers. Also, many applications programs are written by personnel more experienced with the application than with programming techniques, and such a system might significantly improve the quality of their programs.

- b. The question regarding block contents can only be answered in a general way. If we accept the hypothesis that there exists a high degree of commonality in classes of commercial and scientific applications, then a set of blocks can be defined whose members appear frequently in most designs for such classes of applications. Clearly, different blocks are needed for those writing payroll software than for the designers of a wind tunnel simulation. Hence blocks must be grouped in certain sets which form a library. Each library applies to related applications.
- c. The mode of transmittal of the library is a crucial one. A listing of available programs in some volume is of little use to an inexperienced programmer, because such programmer may not know, for example, that he can utilize program #123 to advantage in his design. Just the title of such a program (say Gram-Charlier interpolation) may impact no meaning to our designer. Even if this program has a short description of its use, often such description is for the practitioner rather than for the tyro.

In view of these questions and the stated answers, a technique has been investigated, called here automatic programming. Our implementation of this technique is an interactive terminal system which allows on-line queries.

This work, described in Refs. 3, 4, and 5, consists of two major units, the programs FLOW and AUTO. The program FLOW uses four types of blocks to generate a flowchart of the program. The blocks are:

1. The Control Block. This is a conditional decision block and is written by the programmer.
2. The Functional Block. This is the library block to be used in the design.
3. The Stop Block. This block signals the end of the path and is coded by the STOP statement.
4. The User's Code Block. This block contains the code written by the programmer, exclusively for the application at hand.

Upon completion of the flowchart generation, the control passes to the second unit, that is, to the program AUTO.

The purpose of the program AUTO is to aid the program designer in the choice of the library member. If there are several such suitable library programs, AUTO will help in the selection of the most appropriate one.

How is this done? The user will specify to AUTO what he likes to do, and AUTO will advise which of the available programs do this. AUTO will also tell the characteristics of each method (e.g., will tell that trapezoidal formula approximates each pair

of points of the function by straight line segments) in response to each query.

Further work on this method shall be reported in a future progress report.

Rome Air Development Center  
F30602-74-C-0294

M. L. Shooman and H. Ruston

#### REFERENCES

1. M. Shooman and H. Ruston, "Summary of Technical Progress - Software Modeling Studies," Contract No. F-30602-74-C-0294, Polytechnic Institute of New York, EE/EP74-019, EER 114 (April 1974-September 1974).
2. M. Shooman and H. Ruston, "Summary of Technical Progress - Software Modeling Studies," Contract No. F-30602-74-C-0294, Polytechnic Institute of New York, EE/EP75-004, SMART 100 (October 1974-June 1975).
3. M. Shooman and H. Ruston, "Summary of Technical Progress - Software Modeling Studies," Contract No. F-30602-74-C-0294, Polytechnic Institute of New York, EE/EP76-003, SMART 101 (July 1975-December 1975).
4. M. Shooman and H. Ruston, "Summary of Technical Progress - Software Modeling Studies," Contract No. F-30603-74-C-0294, Polytechnic Institute of New York, EE/EP76-013, SMART 103 (January 1976-June 1976).
5. M. Shooman and H. Ruston, "Summary of Technical Progress - Software Modeling Studies," Contract No. F-30602-74-C-0294, Polytechnic Institute of New York, EE/EP 76-023, SMART 106 (July 1976-December 1976).
6. A. K. Trivedi and M. L. Shooman, "Computer Software Reliability: Many-State Markov Modeling Techniques," Polytechnic Institute of New York, POLY-EE/EP75-005, EER 116 (February 1975).
7. M. Shooman and S. Natarajan, "Effect of Manpower Deployment and Bug Generation on Software Error Models," POLY EE/EP76-007, SMART 102 (May 1976).
8. B. Rudner, "Seeding/Tagging Estimates of the Number of Software Errors: Models and Estimates," POLY EE/EP 76-109, SMART 104 (November 1976).
9. A. Laemmel and M. L. Shooman, "Statistical (Natural) Language Theory and Computer Program Complexity," POLY EE/EP 76-020, SMART 107 (August 1977).
10. A. Laemmel, "Study of Recursive Function Theory and Its Application to Program Complexity," POLY EE77-037 SRS 108 (September 1977).
11. G. S. Popkin, "On the Number of Tests Necessary to Verify a Computer Program," POLY EE77-039, SRS 109 (October 1977).
12. D. Baggi and M. L. Shooman, "An Automatic Driver for Pseudo-Exhaustive Software Testing," submitted for 1978 IEEE Spring Compcon Conference, San Francisco, California (1978).
13. M. L. Shooman, "Software Reliability: Measurement and Models," 1975 Annual Reliability and Maintainability Symposium.
14. M. L. Shooman, "Operational Testing and Software Reliability Estimation During Program Development," 1973 IEEE Symposium on Computer Software Reliability, New York (April 30, 1973).
15. M. L. Shooman, "Probabilistic Models for Software Reliability Prediction," Statistical Methods for the Evaluation of Computer System Performance, Frieberger, Editor, Academic Press, New York (1972).

16. M. L. Shooman, "Structural Models for Software Reliability Prediction, " Second National Conference on Software Reliability, San Francisco (October 1976).
17. H. D. Mills, Internal Memorandum, IBM Federal Systems Division (September 1970).
18. Morton Hyman, personal communication, IBM Federal Systems Division, Morris Plains, New Jersey.
19. M. L. Shooman and S. Sinkar, "Generation of Reliability and Safety Data by Analysis of Expert Opinion, " Proceedings, 1977 Annual Reliability and Maintainability Symposium, Philadelphia, Pa. (Received best paper award.)
20. S. N. Mohanty, "Automatic Program Testing, " Ph.D. Dissertation, Polytechnic Institute of New York (1976).
21. M. L. Shooman, "Software Engineering: Reliability, Design, Management, " CS606 Class Notes, Polytechnic Electrical Engineering Department (Fall 1977).
22. Denis Baggi and M. L. Shooman, "Test Models: Classification and Automatic Driver Design, " POLY EE77-040, SRS 110 (November 1977).
23. A. Laemmel, "Statistical Test Models, " POLY EE77-041, SRS 111 (December 1977).
24. M. Shooman, "Meaning of Exhaustive Software Testing, " PINY EE/EP 74-006, EER 105 (January 1974).
25. M. Shooman and A. Laemmel, "Statistical Theory of Computer Programs - Information Content and Complexity, " Digest of Technical Papers, IEEE Compcon, Fall 1977 (Received best paper award) pp. 341-347.
26. S. N. Mohanty and M. Adamowicz, "Proposed Measures for the Evaluation of Software, " Proceedings of the MRI Symposium on Computer Software Engineering (April 1976).
27. M. H. Halstead, "Software Science, " Elsevier, North-Holland, New York, New York (1977).
28. A. N. Sukert, "A Software Reliability Modeling Study, " RADC-TR-76-24, In-house report (August 1977).
29. A. N. Sukert, "An Investigation of Software Reliability Models, " Proceedings 1977 Annual Reliability and Maintainability Symposium, IEEE, New York, N.Y. p. 478 ff.
30. M. L. Shooman, J. Johnsen, and R. Straub, "Generation of Failure Data Fitting a Specified Hazard Function, " Ninth Reliability and Maintainability Conference, Detroit, Mich., July 1970, Proceedings of Society of Automotive Engineers, Inc., New York, N.Y., pp. 405-413.
31. L. Duval, "Software Data Repository Study, " Report RADC-TR-76-387, Rome Air Development Center Griffiss Air Force Base, New York, 13441 (August 1976).
32. T. J. McCabe, "A Complexity Measure, " IEEE Transactions of Software Engineering, p. 308 (December 1976).
33. C. Marshall, "Applied Graph Theory, " John Wiley, New York (1971).
34. D. Minoli, "Combinatorial Graph Complexity, " Accademia Nazionale Dei Lincei, Estratto dai Rendiconti della Classe di Scienze fisiche, matematiche e naturali, Serie VIII, Vol. LIX (December 1975).
35. R. Flynn, "Dynamic Complexity of Networks, " Ph.D. Dissertation, Department of Mathematics, Polytechnic Institute of Brooklyn (1973).
36. S. Amster, et al, "An Experiment in Automatic Quality Evaluation of Software, " Proceedings of the 1976 Polytechnic Symposium on Computer Software Engineering, John Wiley and Sons, New York (1976).

37. F. P. Brooks, Jr., "The Mythical Man-Month," Addison-Wesley Pub. Co., Reading, Mass., pp. 83-90 (1975).
38. H. H. Goode and R. E. Machol, "System Engineering," McGraw-Hill, New York, 1957, Section 25-2, Group Dynamics, pp. 383-393.
39. G. Weinberg, "The Psychology of Computer Programming," Reinhold-Van Nostrand, New York (1971).
40. M. L. Shooman, "Development of a Structural/Psychological Model of Programmer Productivity and Correctness," Internal memorandum (March 17, 1977).

NOTES ON  $n$ -DIMENSIONAL SYSTEM THEORY

D. C. Youla and G. Gnani

This report makes three observations with regard to several issues of a fundamental nature that apparently must arise in any general theory of linear  $n$ -dimensional systems. It is shown, by means of three specific interrelated counterexamples, that certain decomposition techniques which have proven to be basic for  $n=1$  and  $2$  are no longer applicable for  $n \geq 3$ . In fact, for  $n \geq 3$ , at least three equally meaningful but inequivalent notions of polynomial coprimeness emerge, namely, zero-coprimeness (ZC), minor-coprimeness (MC) and factor-coprimeness (FC). Theorems 1 and 3 clarify the differences (and similarities) between these concepts, and Theorem 2 gives the ZC and MC properties a useful system formulation. (Unfortunately, FC, which in our opinion is destined to play a major role, has thus far eluded the same kind of characterization.) Theorem 4 reveals that the structure of 2-variable elementary polynomial matrices is completely captured by the ZC concept. However, there is reason to believe that ZC is insufficient for  $n \geq 3$  but a counterexample is not at hand. The matter is therefore unresolved.

A. Introduction

This report contains three observations with regard to the structure of  $n$ -dimensional linear systems plus a certain amount of peripheral discussion concerning their implications for System Theory in general.

In the first,  $O_1$ , we show by means of a specific counterexample that the recent important result due to Morf, Levy and Kung<sup>1</sup> on the feasibility of primitive factorization for polynomial matrices in two variables does not generalize to three or more variables ( $n \geq 3$ ).

In the second,  $O_2$ , we explore several possibilities for extending the useful concept of coprimeness of 1-variable polynomial matrices to the  $n$ -dimensional case. It is found that from the point of view of synthesis, the situation changes profoundly for  $n \geq 3$  and that the theory is beset with many of the same difficulties encountered in algebraic geometry.

In the third,  $O_3$ , we illustrate the true importance of the ZC concept by employing it to give a complete characterization of 2-variable elementary polynomial matrices and a partial characterization for  $n \geq 3$ .

$O_1$ . Let  $A(z_1, z_2, \dots, z_n) \equiv A(z)$  denote an  $m \times m$  polynomial matrix in the  $n$  variables  $z_i$ ,  $i = 1 \rightarrow n$ , and let  $d(z) = \det A(z) \neq 0$ .<sup>\*</sup> Suppose that  $d(z) = d_1(z)d_2(z)$  where

---

<sup>\*</sup>  $\det A(z)$  = determinant of  $A(z)$ .

$d_1(z)$  and  $d_2(z)$  are both polynomials. Then, for  $n \geq 3$  it is not always possible to find two polynomial  $m \times m$  matrices  $A_1(z)$  and  $A_2(z)$  such that  $\det A_i(z) = d_i(z)$ ,  $i = 1, 2$  and

$$A(z) = A_1(z)A_2(z) \quad (1)$$

Proof (By counterexample). Let  $\mathcal{C}$  denote the collection of all polynomials  $f(z_1, z_2, z_3)$  in three variables such that  $f(t^3, t^4, t^5) \equiv 0$ . Clearly,  $\mathcal{C}$  contains

$$f_1 = z_2^2 - z_1 z_3 \quad ; \quad f_2 = z_2 z_3 - z_1^3 \quad ; \quad f_3 = z_3^2 - z_1^2 z_2 \quad (2)$$

Actually,  $f_1, f_2$  and  $f_3$  generate  $\mathcal{C}$  in the sense that  $\mathcal{C} = (f_1, f_2, f_3)$ , the set of all linear combinations

$$f(z) = \sum_{i=1}^3 a_i(z) f_i(z) \quad (3)$$

with polynomial coefficients  $a_1(z), a_2(z), a_3(z)$ .

Indeed, any polynomial  $g(z_1, z_2, z_3)$  admits a decomposition

$$g(z_1, z_2, z_3) = z_1^2 a(z_3) + z_1 z_2 b(z_3) + z_1 c(z_3) + z_2 d(z_3) + e(z_3) + f(z_1, z_2, z_3) \quad (4)$$

where  $f \in (f_1, f_2, f_3)$  and  $a, b, c, d$  and  $e$  are polynomials in the single variable  $z_3$ . For observe, that  $g$  is a constant-coefficient sum of monomials of generic type  $z_1^{\ell} z_2^k z_3^r$  and to establish Eq. (4) it is therefore only necessary to establish it for every polynomial  $z_1^{\ell} z_2^k$ . Clearly, the result is true for  $\ell = k = 0$  and we proceed by induction on both  $\ell$  and  $k$ .

Let us write  $g \sim h$  to indicate that  $g(z_1, z_2, z_3)$  is congruent to  $h(z_1, z_2, z_3) \pmod{(f_1, f_2, f_3)^*}$  and suppose that  $z_1^{\ell} z_2^k$  admits the expansion of Equation (4). Then,  $z_1^{\ell+1} z_2^k$  must be congruent to a sum of five polynomials of the type  $z_1^3 a(z_3), z_1^2 z_2 b(z_3), z_1^2 c(z_3), z_1 z_2 d(z_3), z_1 e(z_3)$ . However, from (2),  $z_1^3 \sim z_2 z_3, z_1^2 z_2 \sim z_3^2, z_2^2 \sim z_1 z_3$  and  $z_1^2 z_3 \sim z_2^2 z_1$ . Hence,  $z_1^3 a(z_3) \sim z_2 z_3 a(z_3)$  and  $z_1^2 z_2 b(z_3) \sim z_3^2 b(z_3)$ . Similarly,  $z_1^{\ell} z_2^{k+1}$  is congruent to a sum of five polynomials of the type  $z_1^2 z_2 a(z_3), z_1 z_2^2 b(z_3), z_1 z_2 c(z_3), z_2^2 d(z_3), z_2 e(z_3)$ . But  $z_1^2 z_2 a(z_3) \sim z_3^2 a(z_3), z_1 z_2^2 b(z_3) \sim z_1^2 z_3 b(z_3)$  and  $z_2^2 d(z_3) \sim z_1 z_3 d(z_3)$  so that in both cases the result is again of the form Equation (4). Thus, Eq. (4) is verified.

\* In other words, the difference  $g-h \in (f_1, f_2, f_3)$ .

Now assume that  $g \in \varphi$ . Then,

$$0 \equiv g(t^3, t^4, t^5) = t^6 a(t^5) + t^7 b(t^5) + t^3 c(t^5) + t^4 d(t^5) + e(t^5) \quad (5)$$

Since for arbitrary choices of integers  $m, n, p, q$  and  $r$

$$6 + 5m \neq 7 + 5n \neq 3 + 5p \neq 4 + 5q \neq 5r \quad (6)$$

any cancellation of terms between pairs of polynomials in Eq. (5) is impossible whence,

$$a(t^5) = b(t^5) = c(t^5) = d(t^5) = e(t^5) \equiv 0 \quad .$$

Consequently,  $a(z_3) = b(z_3) = c(z_3) = d(z_3) = e(z_3) \equiv 0$  and  $g(z_1, z_2, z_3) = f(z_1, z_2, z_3) \in (f_1, f_2, f_3)$ , as was to be shown.

By direct calculation,

$$f_2^2 - f_1 f_3 = z_1(z_1^5 - 3z_1^2 z_2 z_3 + z_1 z_2^3 + z_3^3) \quad (7)$$

Let

$$F(z_1, z_2, z_3) = \left[ \begin{array}{c|c} f_2 & f_1 \\ \hline f_3 & f_2 \end{array} \right] \quad (8)$$

and assume that  $F = F_1 F_2$  where  $F_1(z_1, z_2, z_3)$  and  $F_2(z_1, z_2, z_3)$  are  $2 \times 2$  polynomial matrices,

$$\det F_1 = d_1 = z_1 \quad (9)$$

and\*

$$\det F_2 = d_2 = z_1^5 - 3z_1^2 z_2 z_3 + z_1 z_2^3 + z_3^3 \quad (10)$$

Clearly, \*\*  $F(t^3, t^4, t^5) \equiv 0_2$  implies that  $F_2(t^3, t^4, t^5) \equiv 0_2$  since  $\det F_1(t^3, t^4, t^5) = t^3 \neq 0$ . This means that all four entries in  $F_2$  belong to  $\varphi$  from which we conclude that  $d_2 = \det F_2 \in \varphi^2$ . But from what has already been shown,  $\varphi = (f_1, f_2, f_3)$  and it is evident that

$$\varphi^2 = (z_2^2 - z_1 z_3, z_2 z_3 - z_1^3, z_3^2 - z_1^2 z_2)^2 \quad (11)$$

cannot contain any polynomial whose expansion into monomials exhibits a term of degree

\*  $\det F = f_2^2 - f_1 f_3 = d_1 d_2$ .

\*\*  $I_m$  is the  $m \times m$  identity matrix and  $0_{m,r}, 0_m$  denote the  $m \times r$  and  $m \times m$  zero matrices, respectively.

less than four. The factorization  $F = F_1 F_2$ , as described, is therefore impossible, Q.E.D.\*

Comment. The failure of determinantal factorization for  $n \geq 3$  entails some perplexing consequences, at least from the circuit theorist's point of view. As an illustration, consider an  $m \times q$  polynomial matrix  $A(z)$  whose normal rank  $r$  is less than either  $m$  or  $q$ .\*\* Then, in general, it is no longer possible to find two polynomial matrices  $A_1(z)$  and  $A_2(z)$  of respective sizes  $m \times r$  and  $r \times q$  such that

$$A(z) = A_1(z)A_2(z) \quad . \quad (12)$$

In fact, it is easy to construct such an example.

Let

$$A(z) = \left[ \begin{array}{cc|c} f_2 & f_1 & 0 \\ f_3 & f_2 & d_2 \\ \hline 0 & d_1 & f_2 \end{array} \right] = \left[ \begin{array}{c|c} F(z) & \begin{matrix} 0 \\ d_2 \end{matrix} \\ \hline 0 & d_1 \quad f_2 \end{array} \right] \quad (13)$$

where  $f_1, f_2, f_3$  are given in Eq. (2) and  $d_1, d_2$  are defined by Equations (9) and (10). Then,  $\det F = d_1 d_2$  and  $\det A(z) \equiv 0$  so that normal rank  $A(z) = 2$ . Suppose that

$$A(z) = \left[ \begin{array}{c} F_1(z) \\ a \quad b \end{array} \right] \cdot \left[ F_2(z) \middle| \begin{matrix} c \\ d \end{matrix} \right] \quad (14)$$

where  $F_1, F_2$  are  $2 \times 2$  polynomial matrices and  $a(z), b(z), c(z), d(z)$  are four scalar polynomials. From Equations (13) and (14),

$$F = F_1 F_2 \quad ; \quad ac + bd = f_2 \quad (15)$$

According to the counterexample (and the irreducibility of both  $d_1$  and  $d_2$ ), either  $F_1(z)$  or  $F_2(z)$  must be elementary. Let it be  $F_2(z)$ . Then, without loss of generality we can assume that  $F_2(z) = I_2$  and this choice yields  $a(z) \equiv 0$ ,  $b(z) = d_1(z)$  and  $d_1(z)d(z) = f_2(z)$ . But  $d_1(z) = z_1$  does not divide  $f_2(z) = z_2 z_3 - z_1^3$ . Similarly, if  $F_1(z)$  is elementary,  $d_2(z)$  must divide  $f_2(z)$ , another contradiction, Q.E.D.

\*The authors have constructed several other counterexamples but the one given above possesses an appealing amount of structure. (The polynomials  $f_1, f_2, f_3$  have already been used in algebraic geometry to give a concrete example of a prime ideal  $\phi = (f_1, f_2, f_3)$  whose square  $\phi^2$  is not primary.<sup>2</sup> It follows, that in any polynomial decomposition of  $F(z)$  into a product of two square matrices, at least one of the factors must be elementary; i.e., must possess a determinant which is a nonzero constant.

\*\* $A(z)$  has normal rank  $r$  if the largest size minor that does not vanish identically is  $r \times r$ .

### B. On the Coprimeness of Polynomial Matrices

Experience with 1-variable synthesis procedures appears to reinforce the conviction that rank degeneracies can always be eliminated, a priori, by the use of some suitable preamble which extracts superfluous elements. In this process, the availability of a rank-reducing polynomial factorization scheme such as Eq. (12) often plays an important role. However, since such a technique is no longer applicable for  $n \geq 3$ , it is necessary to study the entire matter in greater depth and this we proceed to do in our second observation.

$O_2$ . The subtlety of the  $n$ -dimensional problem becomes quickly apparent when one attempts to frame a notion of polynomial coprimeness which is all-inclusive. A little thought reveals that there are at least three possible meaningful definitions which although equivalent for  $n=1$  and partly equivalent for  $n=2$ , are completely inequivalent for  $n \geq 3$ . (This divergence is undoubtedly responsible for much of the complexity which characterizes the theory of contact of algebraic curves.)

Definitions. Let  $A(z)$  and  $B(z)$  denote, respectively, an  $m \times q$  and  $m \times l$  polynomial matrix,  $q+l \geq m \geq 1$ , and let

$$C(z) = [A(z) | B(z)] \quad (16)$$

Then, 1) the pair  $A(z), B(z)$  is said to be zero left-coprime (ZLC) if there exists no  $n$ -tuple  $z = (z_1, z_2, \dots, z_n)$  which is a zero of all the  $m \times m$  minors of  $C(z)$ , 2) minor left-coprime (MLC) if these minors are relatively prime,\* and 3) factor left-coprime (FLC) if in any polynomial decomposition  $C(z) = C_1(z)C_2(z)$  in which  $C_1(z)$  is square,  $C_1(z)$  is necessarily elementary. (In dual fashion,  $A(z)$  and  $B(z)$  are zero right-coprime (ZRC), etc., if the matrix transposed pair  $A'(z), B'(z)$  is zero left-coprime, etcetera.)

Theorem 1. For  $n=1$  the three definitions are equivalent; i.e.,  $ZLC \equiv MLC \equiv FLC$ . For  $n=2$ ,  $ZLC \neq MLC \equiv FLC$  and for  $n=3$ ,  $ZLC \neq MLC \neq FLC$ . Always,  $ZLC \rightarrow MLC \rightarrow FLC$ .

Proof. That the three definitions are equivalent for  $n=1$  is well-known<sup>4</sup> and is directly attributable to the fact that every ideal of polynomials in one variable is principal.<sup>3</sup> Since the polynomials  $A(z) = z_1$  and  $B(z) = z_2$  possess the common zero  $z = (0, 0)$  but are relatively prime, the ZLC and MLC concepts differ for  $n \geq 2$ . Nevertheless, it is proved in Theorem 3 that for  $n=2$ , a pair  $A(z), B(z)$  is MLC iff it is FLC. For  $n=3$  the situation changes drastically.

\*One or more polynomials are said to be relatively prime if their greatest common polynomial divisor (g.c.d.) is a nonzero constant.

Consider the pair  $A(z) = F(z)$  and  $B(z) = z_1 l_2$  where  $F(z)$  is once again defined by Equation (8). It is easily seen that the greatest common divisor (g.c.d.) of the six  $2 \times 2$  minors of

$$C(z) = [F(z) | z_1 l_2] = \begin{bmatrix} f_2 & f_1 & z_1 & 0 \\ f_3 & f_2 & 0 & z_1 \end{bmatrix} \quad (17)$$

is precisely  $z_1 (=d_1)$ . Thus, the two matrices  $F(z)$  and  $z_1 l_2$  are neither ZLC nor MLC but, as we now show, they are FLC. Indeed, suppose that  $C(z)$  admits a polynomial decomposition

$$C(z) = C_1(z) [C_{21}(z) | C_{22}(z)] \quad (18)$$

where  $C_1$  and  $C_{21}$  are both  $2 \times 2$ . Then,  $F = C_1 C_{21}$  and either  $C_1$  or  $C_{21}$  must be elementary. If  $C_{21}$  is elementary,  $\det C_1 = \text{constant} \cdot \det F = \text{constant} \cdot z_1 d_2$  and from Cauch-Binet it follows that  $d_2 = z_1^5 - 3z_1^2 z_2 z_3 + z_1 z_2^3 + z_3^3$  must divide the g.c.d.  $z_1$ , a contradiction. Hence,  $C_1(z)$  must be elementary and the pair  $F(z), z_1 l_2$  is FLC. In summary, the three properties are inequivalent for  $n \geq 3$ . Furthermore, it is obvious that for all  $n \geq 1$ ,  $ZLC \rightarrow MLC \rightarrow FLC$ , Q.E.D.

**Theorem 2.** 1) the  $m \times q$  and  $m \times \ell$  polynomial matrices  $A(z)$  and  $B(z)$ ,  $q + \ell \geq m \geq 1$ , are ZLC iff there exist two polynomial matrices  $X(z)$  and  $Y(z)$  such that

$$A(z)X(z) + B(z)Y(z) = I_m \quad (19)$$

2) They are MLC iff for every  $i = 1 \rightarrow n$ , there exist polynomial matrices  $X_i(z)$  and  $Y_i(z)$  such that

$$A(z)X_i(z) + B(z)Y_i(z) = \psi_i(z) I_m, \quad (20)$$

where  $\psi_i(z)$  is a nontrivial scalar polynomial independent of the variable  $z_i$ . Moreover, if  $A(z)$  and  $B(z)$  are real,  $X(z)$ ,  $Y(z)$  and  $X_i(z)$ ,  $Y_i(z)$ ,  $\psi_i(z)$ ,  $i = 1 \rightarrow n$ , can always be constructed real.

**Proof.** 1) Clearly, Eq. (19) guarantees that  $\text{rank } C = \text{rank } [A | B] = m$  for all  $z$  and this implies that no  $z = (z_1, z_2, \dots, z_n)$  is a common zero of all the  $m \times m$  minors of  $C(z)$ . Thus Eq. (19) is sufficient for ZLC. To prove necessity we employ a novel technique which succeeds in isolating each individual  $m \times m$  minor of  $C(z)$ .

Let the pair  $A(z)$ ,  $B(z)$  be ZLC and let  $\Delta_{i_1 i_2 \dots i_m}^{i_1 i_2 \dots i_m}(z)$  denote the  $m \times m$  minor of  $C(z)$  formed with the given  $m$  rows and the  $m$  distinct columns numbered  $i_1, i_2, \dots, i_m$ .

From the definition of ZLC, these  $^{q+l}C_m$  polynomials are devoid of any common zeros and invoking a classical result due to Hilbert,<sup>3</sup> there exist polynomials  $a_{i_1 i_2 \dots i_m}(z)$  such that

$$1 = \sum_{(i)} a_{i_1 i_2 \dots i_m}(z) \Delta_{i_1 i_2 \dots i_m}(z) \quad (21)$$

In addition, the  $a$ 's can all be chosen real if all the  $\Delta$ 's are real.

Pick  $K$  to be any  $(q+l) \times m$  real constant matrix whose  $m \times m$  minors  $K \begin{pmatrix} i_1, i_2, \dots, i_m \\ 1, 2, \dots, m \end{pmatrix}$  are all nonzero,<sup>\*</sup> introduce  $q+l$  extra independent variables,  $\lambda_1, \lambda_2, \dots, \lambda_{q+l}$  and let

$$\Lambda(\lambda) = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_{q+l}] \quad (22)$$

The polynomial matrix

$$D(z, \lambda) = C(z) \Lambda(\lambda) K \quad (23)$$

is  $m \times m$  and from the Cauchy-Binet theorem,

$$\Delta(z, \lambda) \equiv \det D(z, \lambda) = \sum_{(i)} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_m} \Delta_{i_1 i_2 \dots i_m}(z) K \begin{pmatrix} i_1, i_2, \dots, i_m \\ 1, 2, \dots, m \end{pmatrix} \quad (24)$$

Thus, for every one of the  $^{q+l}C_m$   $m$ -tuples  $(i) = (i_1, i_2, \dots, i_m)$ ,

$$\Delta_{i_1 i_2 \dots i_m}(z) K \begin{pmatrix} i_1, i_2, \dots, i_m \\ 1, 2, \dots, m \end{pmatrix} = \frac{\partial^m \Delta(z, \lambda)}{\partial \lambda_{i_1} \partial \lambda_{i_2} \dots \partial \lambda_{i_m}} \bigg|_{(\lambda)=0} \quad (25)$$

Let  $D_a(z, \lambda)$  denote the  $m \times m$  polynomial matrix adjugate to  $D(z, \lambda)$ . Since

$$\Delta(z, \lambda) I_m = D(z, \lambda) D_a(z, \lambda) \quad (26)$$

multiplication of both sides of Eq. (23) on the right with  $D_a(z, \lambda)$  yields

$$\Delta(z, \lambda) I_m = C(z) \Lambda(\lambda) K D_a(z, \lambda) \quad (27)$$

In view of Eq. (25), Eq. (27) permits the identifications

$$\Delta_{i_1 i_2 \dots i_m}(z) I_m = C(z) Z_{i_1 i_2 \dots i_m}(z) \quad (28)$$

\* Such  $K$ 's always exist for  $q+l \geq m$ . A diagonal  $k \times k$  matrix with diagonal elements  $a_1, a_2, \dots, a_k$  is written  $\text{diag}[a_1, a_2, \dots, a_k]$ .

where for all (i),

$$Z_{i_1 i_2 \dots i_m}(z) = \frac{1}{K(i_1, i_2, \dots, i_m)} \cdot \frac{\partial^m (\Lambda(\lambda) K D_a(z, \lambda))}{\partial \lambda_{i_1} \partial \lambda_{i_2} \dots \partial \lambda_{i_m}} \bigg|_{(\lambda)=0} \quad (29)$$

is  $(q+l) \times m$  and polynomial. Finally, by combining Eqs. (21) and (28) we reach the desired result Equation (19),

$$C(z)Z(z) = A(z)X(z) + B(z)Y(z) = I_m \quad (30)$$

where

$$Z(z) = \sum_{(i)} a_{i_1 i_2 \dots i_m}(z) Z_{i_1 i_2 \dots i_m}(z) \equiv \begin{bmatrix} X(z) \\ Y(z) \end{bmatrix}. \quad (31)$$

An examination of the above procedure reveals that  $Z(z)$  is always real if  $C(z)$  is real.

2)\* Suppose that the pair  $A(z)$ ,  $B(z)$  satisfies Eq. (20) for every  $i = 1 \rightarrow n$ . Then, by Cauchy-Binet, the g.c.d. of the  $m \times m$  minors of  $C = [A|B]$  must divide every  $\psi_i(z)$ . Since  $\psi_i(z)$  is nontrivial and independent of  $z_i$ ,  $i = 1 \rightarrow n$ , this g.c.d. must be a nonzero constant and the  $^{q+l}C_m \Delta$ 's are therefore relatively prime. Hence,  $A(z)$  and  $B(z)$  are MLC. The necessity of Eq. (20) is also easily established with the aid of Equation (28).

By definition,  $A(z)$  and  $B(z)$  are MLC if the  $\Delta$ 's form a relatively prime set of polynomials. But then, according to another classical result,<sup>3</sup> for every  $i = 1 \rightarrow n$  there exist polynomials  $a_{i_1 i_2 \dots i_m}(z; i)$  such that

$$\sum_{(i)} a_{i_1 i_2 \dots i_m}(z; i) \Delta_{i_1 i_2 \dots i_m}(z) = \psi_i(z) \quad (32)$$

where  $\psi_i(z)$  is nontrivial and independent of  $z_i$ .

As before, the  $a$ 's can be chosen real if all the  $\Delta$ 's are real. Thus, combining Eqs. (28) and (32),

$$C(z)Z_i(z) = A(z)X_i(z) + B(z)Y_i(z) = \psi_i(z)I_m \quad (33)$$

where

$$Z_i(z) = \sum_{(i)} a_{i_1 i_2 \dots i_m}(z; i) Z_{i_1 i_2 \dots i_m}(z) \equiv \begin{bmatrix} X_i(z) \\ Y_i(z) \end{bmatrix}. \quad (34)$$

\* A more constructive approach is given in the proof of the MLP lemma preceding Theorem 4.

The proof of Theorem 2 is complete, Q.E.D.\*

**Corollary.** Let  $A_{11}(z)$ ,  $A_{12}(z)$  and  $A_{21}(z)$  be three polynomial matrices of compatible sizes such that  $A_{21}(z)A_{11}^{-1}(z)A_{12}(z)$  is polynomial. Then, if the pair  $A_{11}, A_{12}$  is MLC,  $A_{21}(z)A_{11}^{-1}(z)$  is polynomial and if the pair  $A_{11}, A_{21}$  is MRC,  $A_{11}^{-1}(z)A_{12}(z)$  is polynomial.

**Proof.** Assume that the pair  $A_{11}, A_{12}$  is MLC. Then, according to Theorem 2, part 2), for every  $i = 1 \rightarrow n$ , there exist polynomial matrices  $X_i(z), Y_i(z)$  such that

$$A_{11}(z)X_i(z) + A_{12}(z)Y_i(z) = \psi_i(z)1 \quad (35)$$

where  $\psi_i(z)$  is nontrivial and independent of  $z_i$ . Thus,

$$A_{21}(z)X_i(z) + A_{21}(z)A_{11}^{-1}(z)A_{12}(z)Y_i(z) = A_{21}(z)A_{11}^{-1}(z)\psi_i(z)1, \quad (36)$$

and from the premise it follows that the product  $A_{21}(z)A_{11}^{-1}(z)\psi_i(z)$  is polynomial for every  $i = 1 \rightarrow n$ . In other words, every denominator of every nontrivial entry in  $A_{21}(z)A_{11}^{-1}(z)$  divides every  $\psi_i(z)$  and is therefore independent of every  $z_i, i = 1 \rightarrow n$ . This is only possible if all such denominators are nonzero constants and  $A_{21}(z)A_{11}^{-1}(z)$  is polynomial, Q.E.D. (A dual argument handles the case  $A_{11}, A_{21}$  MRC.)

Let us return to the problem of factorizing an  $m \times q$  polynomial matrix  $A(z)$  whose normal rank  $r < \min(m, q)$ . By an appropriate interchange of rows and columns we can always assume that

$$A(z) = \left[ \begin{array}{c|c} A_{11}(z) & A_{12}(z) \\ \hline A_{21}(z) & A_{22}(z) \end{array} \right] \begin{matrix} r \\ q-r \\ m-r \end{matrix} \quad (37)$$

where  $\det A_{11}(z) \neq 0$ . Hence, by exploiting a well known matrix result,

$$A_{22}(z) = A_{21}(z)A_{11}^{-1}(z)A_{12}(z) \quad (38)$$

from which it is immediately verified that

$$A(z) = A_1(z)A_2(z) = A_3(z)A_4(z) \quad (39)$$

where

---

\* The method employed in the proof of Theorem 2 can also be used in the 1-variable case and avoids any appeal to the Smith canonic form.

$$A_1(z) = \begin{bmatrix} I_r \\ A_{21}A_{11}^{-1} \end{bmatrix}; \quad A_3(z) = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} \quad (40)$$

and

$$A_2(z) = [A_{11} | A_{12}]; \quad A_4(z) = [I_r | A_{11}^{-1}A_{12}]. \quad (41)$$

According to the corollary to Theorem 2, if the pair  $A_{11}, A_{12}$  is MLC or if the pair  $A_{11}, A_{21}$  is MRC, Eq. (38) is sufficient to assure that at least one of the two breakdowns in Eq. (39) is completely polynomial. For  $n=1$  and 2 this minor-coprimeness can always be accomplished by either extracting a square polynomial factor from the left of  $A_2(z)$  or from the right of  $A_3(z)$  and using the polynomial quotients, but for  $n \geq 3$  the inequivalence of MC and FC precludes any such general possibility.

Theorem 3. For  $n=2$ , a polynomial pair  $A(z), B(z)$  is MLC iff it is FLC.

Proof (Necessity). Let the pair  $A(z), B(z)$  be MLC but not FLC. Then,  $C(z) = [A(z) | B(z)]$  admits a polynomial decomposition  $C(z) = C_1(z)C_2(z)$  wherein  $C_1(z)$  is square and non-elementary. Since MLC implies that normal rank  $C(z) = m$ ,  $\det C_1(z)$  is a nonconstant polynomial which divides all the  $m \times m$  minors of  $C(z)$ , a contradiction.

Sufficiency. Let  $A(z)$  and  $B(z)$  be FLC and assume first that normal rank  $C(z) = m$ . Let the g.c.d.  $\psi(z)$  of the  $m \times m$  minors of  $C(z)$  possess a linear divisor  $z_1 - \eta$ ,  $\eta$  a constant. Then, normal rank  $C(\eta, z_2) < m$  and there exists an  $m \times 1$  polynomial column-vector  $\underline{x}(z_2)$  in the single variable  $z_2$  such that\*

$$\underline{x}'(z_2)C(\eta, z_2) \equiv \underline{0}'_{q+l} \quad (42)$$

Obviously, there is no loss of generality in assuming that the  $m$  components of  $\underline{x}(z_2)$  are relatively prime.

Thus,  $\underline{x}(z_2)$  can be incorporated into the first column of an  $m \times m$  elementary polynomial matrix  $X(z_2)$  and it follows from Eq. (42) that the first row of

$$D(z) = X'(z_2)C(z) \quad (43)$$

vanishes identically in  $z_2$  for  $z_1 = \eta$ . Consequently, if\*\*

\* Column-vectors are written  $\underline{a}, \underline{x}$ , etc., and  $A'$  denotes the matrix transpose of  $A$ .

\*\*  $A \dot{+} B$  is the "direct" sum of  $A$  and  $B$ .

$$\Lambda(z_1) = (z_1 - \eta)1_1 + 1_{m-1} \quad , \quad (44)$$

the matrix  $C_2(z) = \Lambda^{-1}(z_1)D(z)$  is polynomial as is the decomposition

$$C(z) = (X'(z_2))^{-1} \Lambda(z_1) C_2(z) = C_1(z) C_2(z) \quad . \quad (45)$$

But

$$\det C_1(z) = \text{constant} \cdot (z_1 - \eta) \quad , \quad (46)$$

which is incompatible with the FLC assumption. Similarly,  $\psi(z)$  can possess no linear divisors of the form  $z_2 - \xi$ ,  $\xi$  a constant. In short, any irreducible factor of  $\psi(z)$  must be a 2-variable polynomial. (It is also important to observe that if  $C(z)$  is square, the factor  $C_1(z)$  in Eq. (45) can be extracted from the right as well as from the left.)

Since the entries in  $C(z)$  can be viewed as polynomials in  $z_2$  with coefficients that are polynomials in  $z_1$ , it is easily shown<sup>1</sup> that there exists a square polynomial matrix  $V(z)$  whose determinant is a nontrivial polynomial in the single variable  $z_1$ , such that

$$C(z)V(z) = [E(z)]_{0, q+l-m}^{m, q+l-m} \quad . \quad (47)$$

By Cauchy-Binet,  $\psi(z)$  divides  $\det E(z)$ . With the aid of the factorization technique described above, it is possible to effect a polynomial decomposition

$$E(z) = C_1(z)F(z) \quad (48)$$

in which  $\det C_1(z)$  is devoid of any linear divisors of the form  $z_1 - \eta$  and  $\det F(z)$  is a nontrivial polynomial in the single variable  $z_1$ . Clearly then, because  $\psi(z)$  is free of all 1-variable divisors, it must divide  $\det C_1(z)$ .

Thus,

$$C(z) = C_1(z)[F(z)]_{0, q+l-m}^{m, q+l-m} V^{-1}(z) = C_1(z)C_2(z) \quad . \quad (49)$$

The matrix

$$C_2(z) = [F(z)]_{0, q+l-m}^{m, q+l-m} V^{-1}(z) \quad (50)$$

is polynomial.\* In fact,  $C_2(z) = M(z)/g(z_1)$  where  $M(z)$  is polynomial and  $g(z_1)$  is a nontrivial scalar polynomial in the single variable  $z_1$ . Since  $C(z)$  is polynomial, any

\* This key observation is a special case of lemma 2.1, Reference 1.

linear factor  $z_1 - \eta$  of  $g(z_1)$  must divide the product  $C_1(z)M(z)$ ; i. e.,

$$C_1(\eta, z_2)M(\eta, z_2) \equiv 0_{m, q+l} \quad (51)$$

But, being free of the divisor  $z_1 - \eta$ ,  $\det C_1(\eta, z_2) \neq 0$  whence,

$$M(\eta, z_2) \equiv 0_{m, q+l} \quad (52)$$

and the polynomial  $z_1 - \eta$  divides  $M(z_1, z_2)$ . Since all such linear factors can be cancelled by a repetition of this cycle,  $g(z_1)$  itself must divide  $M(z)$  and  $C_2(z)$  is therefore polynomial, as stated.

By invoking the FLC property it is now concluded that  $C_1(z)$  is elementary,  $\psi(z)$  is a nonzero constant and the pair  $A(z), B(z)$  is MLC. Finally, to complete the proof of Theorem 3 it must be shown that the FLC hypothesis excludes the possibility that normal rank  $C(z) < m$ .

Let normal rank  $C(z) = r < m$ . Then, for some choice of square constant permutation matrices  $P$  and  $Q$ ,

$$PC(z)Q = \left[ \begin{array}{c|c} r & q+l-r \\ \hline A_{11}(z) & A_{12}(z) \\ A_{21}(z) & A_{22}(z) \end{array} \right] \begin{array}{l} r \\ m-r \end{array} \quad (53)$$

where  $\det A_{11}(z) \neq 0$  and  $A_{22}(z) = A_{21}(z)A_{11}^{-1}(z)A_{12}(z)$ . It should be clear that  $A(z), B(z)$  FLC implies, a fortiori, that the pair  $A_{11}(z), A_{12}(z)$  is also FLC. But then, from what has already been proved, since the  $rx(q+l)$  matrix

$$A_2(z) = [A_{11}(z) | A_{12}(z)] \quad (54)$$

has normal rank  $r$ ,  $A_{11}(z)$  and  $A_{12}(z)$  must be MLC and according to the corollary to Theorem 2,  $A_{21}(z)A_{11}^{-1}(z)$  is polynomial. Thus, the factorization

$$C(z) = P^{-1} \left[ \begin{array}{c|c} I_r & \\ \hline A_{21}A_{11}^{-1} & 0_{m, m-r} \end{array} \right] \cdot \left[ \begin{array}{c} A_2 \\ 0_{m-r, q+l} \end{array} \right] Q^{-1} = C_1(z) \cdot C_2(z) \quad (55)$$

is polynomial but  $C_1(z)$  is square and non-elementary, a contradiction, Q. E. D.

(There is no doubt that the equivalence of MLC and FLC for  $n=2$  goes a long way towards explaining why certain 1-variable synthesis ideas admit reasonably successful 2-dimensional generalizations.)

$O_3$ . Imbedded in the sufficiency part of the proof of Theorem 3 is a simple algorithm for the removal of any 1-variable divisor of the g.c.d.  $\psi(z_1, z_2)$  by the extraction of appropriate matrix-polynomial factors. It is distinguished mainly by its explicit use of the known result that any polynomial row-vector in one variable whose components are relatively prime can be incorporated into the first row of an elementary polynomial matrix. This strongly suggests that an understanding of the structure of  $n$ -dimensional elementary polynomial matrices can provide insight and guidance in the formulation and solution of the general synthesis problem. Theorem 4 gives a complete answer for  $n=2$  and a partial answer for  $n \geq 3$  with the aid of the zero left-prime concept. It also makes essential use of an auxiliary lemma of considerable interest in its own right.

The MLP lemma. \* An  $m \times r$  polynomial matrix  $C(z)$ ,  $m \leq r$ , is MLP iff for every  $i = 1 \rightarrow n$ , it can be incorporated into the first  $m$  rows of a polynomial  $r \times r$  matrix  $C_e(z)$  whose determinant  $\psi_i(z)$  is nontrivial and independent of  $z_i$ . Moreover,  $C_e(z)$  can be constructed real if  $C(z)$  is real.

Proof. (Sufficiency). Suppose that such an enlargement is possible for every  $i = 1 \rightarrow n$ . Then, from the adjugate identity

$$C_e(z)(C_e(z))_a = \psi_i(z)1_r \quad (56)$$

we obtain, for every  $i = 1 \rightarrow n$ , an  $r \times m$  polynomial matrix  $Z_i(z)$  which satisfies

$$C(z)Z_i(z) = \psi_i(z)1_m \quad (57)$$

and the MLP character of  $C(z)$  follows by exactly the same reasoning employed in part 2) of Theorem 2.

Necessity. For every fixed  $i$ ,  $1 \leq i \leq n$ , the entries in  $C(z)$  can be viewed as polynomials in  $z_i$  with coefficients that are polynomials in the remaining  $n-1$  variables. Hence, by adapting the standard 1-variable argument used to derive the Smith form, it can be asserted that there exists an  $r \times r$  polynomial matrix  $V(z)$  whose determinant  $d(z)$  is nontrivial and independent of  $z_i$  such that, \*\*

$$C(z)V(z) = [E(z) | 0_{m, r-m}] \quad (58)$$

Let  $C_1(z)$  denote the matrix formed with the last  $r-m$  rows of the adjugate of  $V(z)$ .

\* ZLP, MLP and FLP abbreviate zero left-prime, minor left-prime and factor left-prime, respectively. Their meanings are of course identical with those given in the previous definitions for ZLC, MLC and FLC of  $C(z) = [A(z) | B(z)]$ .

\*\*  $V(z)$  and  $E(z)$  can be constructed real if  $C(z)$  is real.

Clearly,

$$C_e(z) = \begin{bmatrix} C(z) \\ C_1(z) \end{bmatrix} \quad (59)$$

incorporates  $C(z)$  into its first  $m$  rows and since

$$C_1(z)V(z) = [0_{r-m,m} \mid d(z)1_{r-m}] \quad (60)$$

Equation (58) yields

$$C_e(z)V(z) = \left[ \begin{array}{c|c} E(z) & 0_{m,r-m} \\ \hline 0_{r-m,m} & d(z)1_{r-m} \end{array} \right] \quad (61)$$

Thus,

$$\det C_e(z) = d^{r-m-1}(z) \cdot \det E(z) \quad (62)$$

is independent of  $z_i$  iff  $\det E(z)$  is independent of  $z_i$ .

From Eq. (58),

$$d(z)C(z) = E(z) \begin{bmatrix} 1_m & 0_{m,r-m} \end{bmatrix} V_a(z) \quad (63)$$

and by Cauchy-Binet,  $\det E(z)$  divides the g.c.d. of the  $m \times m$  minors of the left-hand side of Equation (63). By hypothesis  $C(z)$  is MLP and this g.c.d. equals  $d^m(z)$ . Consequently,  $\det E(z)$  and therefore  $\det C_e(z)$ , are both independent of  $z_i$ , Q.E.D.

**Theorem 4.** Let  $C(z)$  denote an  $m \times r$  polynomial matrix,  $m \leq r$ . Then, 1) in order that  $C(z)$  be incorporable into the first  $m$  rows of an  $r \times r$  elementary polynomial matrix  $U_e(z)$  it is necessary that it be ZLP.

2) For  $n=2$  the ZLP condition is also sufficient.

3) For  $n \geq 3$  the ZLP condition is sufficient if it can be proved to suffice for  $m=1$ .

**Proof.** 1) Clearly, any common zero of all the  $m \times m$  minors of  $C(z)$  must appear as a zero of the determinant of any  $r \times r$  polynomial matrix which includes  $C(z)$  in  $m$  of its rows. Thus, no such matrix can be elementary and the ZLP property is therefore necessary.

2) Let  $C(z)$  be ZLP and suppose that  $n=2$ . Then, from the ZLP property alone, it follows from Theorem 2, Part 1), that there exists a 2-variable  $r \times m$  polynomial matrix  $Z(z)$  such that

$$C(z)Z(z) = I_m \quad (64)$$

Evidently Eq. (64) implies that  $C(z)$  and  $Z(z)$  are MLP and MRP, respectively.

Thus, invoking the MLP lemma (and its dual), there exist polynomial matrices  $C_1(z)$  and  $Z_1(z)$  of sizes  $(r-m) \times m$  and  $r \times (r-m)$  such that the determinants of the two  $r \times r$  matrices

$$C_e(z) = \left[ \begin{array}{c|c} C(z) & \\ \hline C_1(z) & \end{array} \right], \quad Z_e(z) = [Z(z) | Z_1(z)] \quad (65)$$

are independent of  $z_1$  and  $z_2$ , respectively; i.e.,

$$\det C_e(z) = f(z_2), \quad \det Z_e(z) = g(z_1) \quad (66)$$

By direct multiplication,

$$C_e(z)Z_e(z) = \left[ \begin{array}{c|c} I_m & CZ_1 \\ \hline C_1Z & C_1Z_1 \end{array} \right]; \quad (67)$$

or, after performing some obvious row and column operations which wipe out  $C_1Z$  and  $CZ_1$  (and leave determinants unchanged),

$$\left[ \begin{array}{c} C \\ \hline C_1(1_r - ZC) \end{array} \right] \cdot [Z | 1_r - ZC] Z_1 = I_m + C_1(1_r - ZC)Z_1 \quad (68)$$

Let  $L(z) = C_1(1_r - ZC)Z_1$ . It follows from (68) that  $\det L(z) = f(z_2)g(z_1)$ , a product of two 1-variable polynomials! It is therefore possible to employ the factorization algorithm for the extraction of 1-variable determinantal factors (described in the proof of Theorem 3) to effect a decomposition  $L(z) = L_2(z)L_1(z)$  where  $L_2(z)$  and  $L_1(z)$  are both  $(r-m) \times (r-m)$ ,  $\det L_2(z) = f(z_2)$  and  $\det L_1(z) = g(z_1)$ . By the same argument used to establish the polynomiality of  $C_2(z)$ , Eq. (50), Theorem 3, it follows that  $Z_2 = (1_r - ZC)Z_1L_1^{-1}$  and  $C_2 = L_2^{-1}C_1(1_r - ZC)$  are both polynomial. But then, Eq. (68) gives

$$\left[ \begin{array}{c} C(z) \\ \hline C_2(z) \end{array} \right] \cdot [Z(z) | Z_2(z)] = I_r \quad (69)$$

and the  $r \times r$  matrix

$$U_e(z) = \left[ \begin{array}{c} C(z) \\ \hline C_2(z) \end{array} \right] \quad (70)$$

incorporates  $C(z)$  into its first  $m$  rows and is also elementary.

3) It is the authors' belief that ZLP is no longer sufficient for  $n \geq 3$  but a counter-example is not at hand. Nevertheless, as a positive contribution we have developed a simple recursive method which proceeds by induction and shows that all the difficulty resides in the case  $m = 1$ .

Suppose therefore that ZLP is sufficient for all  $m \leq s$ . We prove that it also suffices for  $m = s+1$ . Let the  $(s+1) \times r$  polynomial matrix  $C(z)$ ,  $s+1 \leq r$ , possess the ZLP property. Then, a fortiori, the  $s \times r$  polynomial matrix  $C_1(z)$  formed with the first  $s$  rows of  $C(z)$  is also ZLP and by the induction hypothesis there exists an  $r \times r$  elementary polynomial matrix  $F(z)$  which incorporates  $C_1(z)$  into its first  $s$  rows.

Let  $\underline{c}'(z)$  denote the last row of  $C(z)$ . Clearly, the solution  $\underline{x}'(z)$  of the equation

$$\underline{x}'(z)F(z) = \underline{c}'(z) \quad (71)$$

is polynomial. Write  $\underline{x}' = [\underline{x}'_1 | \underline{x}'_2]$  where  $\underline{x}'_1$  is  $1 \times s$  and  $\underline{x}'_2$  is  $1 \times (r-s)$ . Then, because of the mode of construction of  $F(z)$  and the choice of  $\underline{x}'(z)$ ,

$$\left[ \begin{array}{c|c} 1_s & 0_{s, r-s} \\ \hline \underline{x}'_1(z) & \underline{x}'_2(z) \end{array} \right] \cdot F(z) = C(z) \quad (72)$$

It is not difficult to see from Eq. (72) that any common zero of the  $r-s$  polynomial components of  $\underline{x}'_2(z)$  is necessarily a zero of all the  $(s+1) \times (s+1)$  minors of  $C(z)$ . Hence,  $\underline{x}'_2(z)$  must be ZLP.

By the induction hypothesis,  $\underline{x}'_2(z)$  can be incorporated into the first row of an  $(r-s) \times (r-s)$  elementary polynomial matrix  $X(z)$  and it is easily seen that

$$U_e(z) = \left[ \begin{array}{c|c} 1_s & 0_{s, r-s} \\ \hline \underline{x}'_1(z) & X(z) \\ \hline \underline{Y}(z) & \end{array} \right] \cdot F(z) \quad (73)$$

is an  $r \times r$  polynomial matrix which incorporates  $C(z)$  into its first  $s+1$  rows, irrespective of the choice of  $(r-s-1) \times s$  polynomial matrix  $Y(z)$ . But

$$\det U_e(z) = \det X(z) \cdot \det F(z) = \text{nonzero constant} \quad (74)$$

so that  $U_e(z)$  is elementary and the proof of Theorem 4 is complete, Q.E.D.

Final comment. Theorems 1 to 4 have the correct system emphasis and should prove useful, but unfortunately, the authors have as yet been unable to translate the FC property into a tractable criterion. This is a serious hiatus, because in our

opinion it is FC which lies at the heart of the  $n$ -dimensional synthesis problem. We hope to address this and other related open questions in a future paper.

Rome Air Development Center  
F-30602-78-C-0048

D. C. Youla and G. Gnani

#### REFERENCES

1. M. B. Morf, B. Levy and S. Y. Kung, "New Results in 2-D System Theory, Part I: 2-D Polynomial Matrices, Factorization and Coprimeness," Proc. IEEE, pp. 861-872 (June 1977).
2. D. G. Northcott, "Ideal Theory," Cambridge at the University Press (1953).
3. Van der Waerden, "Modern Algebra," Vol. II, Frederick Ungar Pub. Co., N. Y. (1950).
4. H. H. Rosenbrock, "State-Space and Multivariable Theory," Wiley, N. Y. (1970).

# THE IDENTIFICATION OF LINEAR DYNAMICAL SYSTEMS FROM TIME-DOMAIN MEASUREMENTS: A CRITICAL STUDY OF CERTAIN ASPECTS OF PRONY'S METHOD

D. C. Youla

The problem of identifying the impulse response of a lumped, linear time-invariant system from a discrete set of its samples, is essentially finite-dimensional in character.

Prony's method is distinguished by the fact that it recognizes this finite-dimensionality from the outset and thereby succeeds in avoiding much of the numerical instability exhibited by algorithms that are designed to work in an infinite-dimensional setting.

This report attempts to generate some new insight into both the theoretical and practical implications of Prony's ideas and it is believed that Theorems 1-4 make a modest contribution in this direction.

## A. Formulation

Let

$$H(s) = \frac{b_0 + b_1 s + \dots + b_{n-1} s^{n-1}}{a_0 + a_1 s + \dots + a_n s^n}, \quad a_n \neq 0, \quad (1)$$

denote the transfer function of a dynamical causal linear time-invariant analogue system. Let  $s_1, s_2, \dots, s_q$  denote the distinct poles of  $H(s)$  and  $m_1, m_2, \dots, m_q$  their respective multiplicities. Then, the associated impulse response  $h(t) \approx H(s)$  and is of the form

$$h(t) = \sum_{i=1}^q P_i(t) e^{s_i t}, \quad t > 0, \quad (2)$$

where  $P_i(t)$  is a polynomial of degree  $m_i - 1$ ,  $i = 1 \rightarrow q$ , and

$$m_1 + m_2 + \dots + m_q = n. \quad (3)$$

Of course, the  $s_i$ 's occur in complex-conjugate pairs and the defining coefficients in  $H(s)$  and in all  $P$ 's are real. It is easily shown from Eq. (2) that

$$H(s) = \sum_{i=1}^q \left( P_i \left( \frac{\partial}{\partial x} \right) \frac{1}{s-x} \right)_{x=s_i}. \quad (4)$$

By definition, the z-transform of the 1-sided sequence

$$c_k = h(kT), \quad k = 0 \rightarrow \infty, \quad (5)$$

obtained by sampling  $h(t)$  once every  $T$  seconds is given by

$$W(z) = \sum_{k=0}^{\infty} c_k z^{-k} = \sum_{k=0}^{\infty} h(kT) z^{-k} \quad (6)$$

$$= \sum_{k=0}^{\infty} z^{-k} \sum_{i=1}^q P_i(kT) e^{s_i kT} \quad (7)$$

$$= \sum_{i=1}^q \sum_{k=0}^{\infty} P_i(kT) e^{s_i kT} z^{-k} \quad (8)$$

$$= \sum_{i=1}^q \sum_{k=0}^{\infty} \left( P_i \left( \frac{\partial}{\partial s} \right) e^{skT} \right)_{s=s_i} \cdot z^{-k} \quad (9)$$

$$= \sum_{i=1}^q \left( P_i \left( \frac{\partial}{\partial s} \right) \sum_{k=0}^{\infty} e^{skT} z^{-k} \right)_{s=s_i} \quad (10)$$

$$= \sum_{i=1}^q \left( P_i \left( \frac{\partial}{\partial s} \right) \frac{1}{1 - e^{sT} z^{-1}} \right)_{s=s_i} \quad (11)$$

In particular, if the poles of the analogue system are all simple, each  $P_i(t)$  reduces to a real constant  $A_i$ ,  $i = 1 \rightarrow n$ , and Eq. (11) collapses into the familiar formula,<sup>1</sup>

$$W(z) = \sum_{i=1}^n \frac{A_i}{1 - e^{s_i T} z^{-1}} \quad (12)$$

Observe that

- (1)  $W(z)$  is analytic in a neighborhood of  $z = \infty$  (Causality),
- (2) vanishes at  $z = 0$  ( $h(t)$  is free of impulses) and
- (3) possesses the distinct poles

$$z_i = e^{s_i T} \quad (13)$$

with respective multiplicities  $m_i$ ,  $i = 1 \rightarrow q$ .<sup>(a)</sup>

<sup>(a)</sup> Under the transformation  $z = e^{sT}$ , two  $s_i$ 's with the same real parts but with imaginary parts that differ by an integer multiple of  $2\pi/T$  map into the same  $z_i$ . We assume that the sampling rate  $1/T$  is high enough to preclude any such "folding" effect.

In short, the task of identifying  $h(t)$  from a knowledge of  $\ell + 1$  of its samples

$$c_k = h(kT), \quad k = 0 \rightarrow \ell, \quad (14)$$

has been reduced to the following purely algebraic problem: 1) Given  $\ell + 1$  real numbers  $c_0, c_1, \dots, c_\ell$  find a real rational function  $W(z)$  which is expressible in the irreducible form

$$W(z) = \frac{z(f_0 + f_1 z + \dots + f_{n-1} z^{n-1})}{1 + g_1 z + g_2 z^2 + \dots + g_n z^n} \equiv \frac{zf(z)}{g(z)}, \quad (15)$$

$$n \geq 1; \quad g_n \neq 0$$

and whose Laurent expansion

$$W(z) = c_0 + \frac{c_1}{z} + \frac{c_2}{z^2} + \dots + \frac{c_\ell}{z^\ell} + \sum_{k=\ell+1}^{\infty} c_k z^{-k} \quad (16)$$

about  $z = \infty$  possesses the prescribed coefficients  $c_k$ ,  $k = 0 \rightarrow \ell$ . (By irreducible we mean that  $f(z)$  and  $g(z)$  are relatively prime.)

2) Having determined  $W(z)$  in step 1), perform the partial fraction expansion

$$\frac{W(z)}{z} = \sum_{i=1}^q \sum_{k=1}^{m_i} B_k^{(i)} (z - z_i)^{-k} \quad (16a)$$

and use the  $B$ 's to calculate the  $A$ 's in the partial fraction expansion,

$$H(s) = \sum_{i=1}^q \sum_{k=1}^{m_i} A_k^{(i)} (s - s_i)^{-k}. \quad (16b)$$

If all poles are simple, every  $m_i$  is equal to unity and we can immediately make the complete identifications

$$A_1^{(i)} = B_1^{(i)}, \quad i = 1 \rightarrow n. \quad (16c)$$

However, if some of the poles are multiple, the relationship between the  $B$ 's and  $A$ 's is rather subtle and it becomes necessary to exploit the interconnection formulas (4) and (11). (Another approach is described in Theorem 4.)

Step 1) is the subject of Theorems 1 and 2 and the answer to 2) is contained in Theorem 3.

## B. The Main Results

From Eqs. (15) and (16),

$$z(f_0 + f_1 z + \dots + f_{n-1} z^{n-1}) = (1 + g_1 z + \dots + g_n z^n) \cdot \sum_{k=0}^{\infty} c_k z^{-k} \quad (17)$$

and equating coefficients of like powers of  $z$  on both sides we get

$$\begin{aligned} f_{n-1} &= g_n c_0 \\ f_{n-2} &= g_{n-1} c_0 + g_n c_1 \\ &\vdots \\ &\vdots \end{aligned} \quad (18)$$

$$f_0 = g_1 c_0 + g_2 c_1 + \dots + g_n c_{n-1}, \quad (18)$$

$$0 = c_r + g_1 c_{r+1} + g_2 c_{r+2} + \dots + g_n c_{r+n}, \quad r \geq 0. \quad (19)$$

The  $n+1$  equations in Eq. (19) obtained by letting  $r$  range from 0 to  $n$  can be written in the matrix form

$$\begin{bmatrix} c_0 & c_1 & \dots & c_n \\ c_1 & c_2 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ c_n & c_{n+1} & \dots & c_{2n} \end{bmatrix} \begin{bmatrix} 1 \\ g_1 \\ \vdots \\ g_{n-1} \\ g_n \end{bmatrix} = \underline{0}_{n+1}. \quad (20)$$

Or, in an obvious notation,

$$C_n \underline{\zeta} = \underline{0}_{n+1} \quad (21)$$

where  $C_n$  is the coefficient matrix and

$$\underline{\zeta} = \begin{bmatrix} 1 \\ g \end{bmatrix}; \quad g = (g_1, g_2, \dots, g_n)' \quad (22)$$

A solution

$$\underline{\zeta} = (g_0, g_1, \dots, g_n)' \quad (23)$$

of Eq. (21) is said to be admissible if

$$1) \ g_0 g_n \neq 0 \text{ and} \quad (24)$$

2) if the polynomial  $f(z)$  generated by  $\underline{\zeta}$  by means of the recursions contained in Eq. (18) is relatively prime to

$$g(z) = g_0 + g_1 z + \dots + g_n z^n. \quad (25)$$

Thus, the rational function  $W(z) = zf(z)/g(z)$  defined by an admissible  $\underline{\zeta}$  is of degree precisely equal to  $n$ . Theorem 1 gives an exact and useful characterization of admissibility.

Theorem 1. Let

$$C_{n-1} = \begin{bmatrix} c_0 & c_1 & \dots & c_{n-1} \\ c_1 & c_2 & \dots & c_n \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-2} \end{bmatrix}; \quad H_n = \begin{bmatrix} c_1 & c_2 & \dots & c_n \\ c_2 & c_3 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_n & c_{n+1} & \dots & c_{2n} \end{bmatrix} \quad (26)$$

Then, the homogeneous linear system Eq. (21) possesses an admissible solution  $\underline{\zeta}$  iff

$$\text{rank } C_n = \text{rank } C_{n-1} = \text{rank } H_n = n. \quad (27)$$

Such a solution is unique up to multiplication by a nonzero scalar.

Proof. Sufficiency. Suppose that Eq. (27) is true. Then,  $C_n$  is singular and Eq. (21) possesses one linearly independent solution,  $\underline{\zeta} = (g_0, g_1, \dots, g_n)'$ . We maintain that  $g_0 g_n \neq 0$ . Indeed,  $g_n = 0$  implies that

$$C_{n-1} \begin{bmatrix} g_0 \\ g_1 \\ \cdot \\ \cdot \\ g_{n-1} \end{bmatrix} = \underline{0}_n \quad (28)$$

while  $g_0 = 0$  implies that

$$H_n \begin{bmatrix} g_1 \\ g_2 \\ \cdot \\ \cdot \\ g_n \end{bmatrix} = \underline{0}_n \quad (29)$$

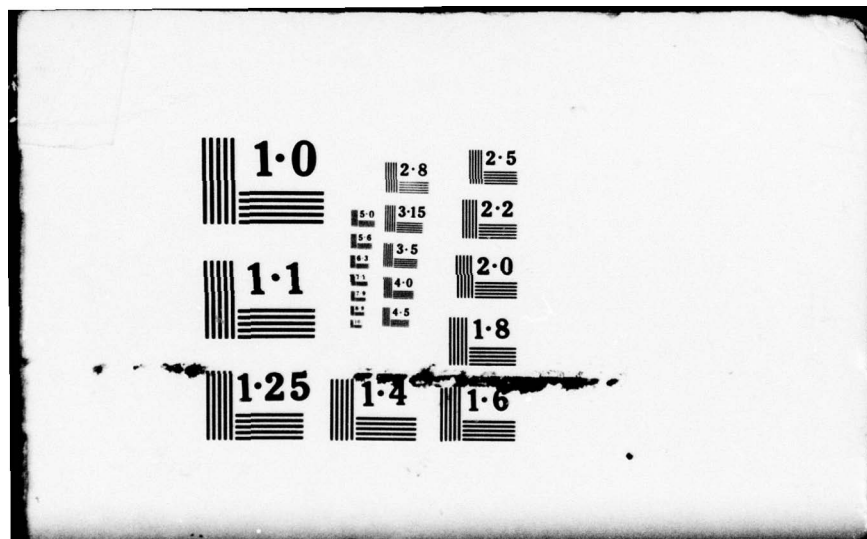
AD-A063 181

POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 9/3  
PROGRESS REPORT NUMBER 43 TO THE JOINT SERVICES TECHNICAL ADVIS--ETC(U)  
NOV 78 A A OLINER F44620-78-C-0074  
POLY-MRI-452.43-78 NL

UNCLASSIFIED

6 OF 6  
AD A  
063181





1.0

2.8

2.5

5.0

3.15

2.2

5.6

3.5

2.0

6.3

4.0

7.1

4.5

1.8

1.1

1.25

1.4

1.6

and in both cases the nonsingularity of  $C_{n-1}$  and  $H_n$ , respectively, forces  $\underline{c}$  to be the zero vector, a contradiction. Hence, without loss of generality we may set  $g_0 = 1$  and  $g(z) = 1 + g_1 z + \dots + g_n z^n$ ,  $g_n \neq 0$ .

By construction,  $g(z)$  and the polynomial  $f(z)$  generated by  $\underline{c}$  with the aid of Eq. (17), define a rational function  $W(z) = zf(z)/g(z)$  whose Laurent expansion about  $z = \infty$  possesses  $c_0, c_1, \dots, c_{2n}$  as its first  $2n+1$  coefficients. This implies that  $f(z)$  is relatively prime to  $g(z)$ . Suppose the contrary. Then,  $\text{degree } W(z) \leq n-1$  and from a well known result,  $\text{rank } H_n \leq n-1$ , a contradiction.<sup>(b)</sup> Thus,  $W(z)$  has all the desired properties.

Necessity. Let Eq. (21) possess an admissible solution  $\underline{c} = (1, g_1, \dots, g_n)'$  and let  $f(z)$  denote the associated polynomial synthesized by the recursive scheme Equation (17). From the definition of admissibility,  $W(z) = zf(z)/g(z)$  has degree  $n$ . Moreover, since it is again obvious that the Laurent expansion of  $W(z)$  about  $z = \infty$  duplicates the coefficients  $c_0, c_1, \dots, c_{2n}$  correctly,  $\text{rank } H_n = n$ .

Now clearly, the function  $W(z) = f(z)/g(z)$  is also of degree  $n$  and possesses the Laurent expansion

$$\frac{W(z)}{z} = \sum_{k=0}^{\infty} c_k z^{-(k+1)} \quad (30)$$

Thus, appealing once again to the result quoted in the footnote below, it is concluded that  $\text{rank } C_n = \text{rank } C_{n-1} = n = \text{rank } H_n$ , Q. E. D.

Corollary. Under the conditions of Theorem 1,

$$(-1)^{n-1} \frac{W(z)}{z} = \begin{vmatrix} c_0 & c_1 & \dots & c_n \\ c_1 & c_2 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ 0 & \psi_0(z) & \dots & \psi_{n-1}(z) \end{vmatrix} \div \begin{vmatrix} c_0 & c_1 & \dots & c_n \\ c_1 & c_2 & \dots & c_{n+1} \\ \dots & \dots & \dots & \dots \\ c_{n-1} & c_n & \dots & c_{2n-1} \\ 1 & z & \dots & z^n \end{vmatrix} \quad (31)$$

<sup>(b)</sup> Consider any rational function  $W(z)$  of degree  $\leq n$  whose Laurent expansion at  $z = \infty$  is given by

$$W(z) = \sum_{k=0}^{\infty} c_k z^{-k}.$$

Then, <sup>2</sup>  $\text{degree } W(z) = \text{rank } H_{n+r}$ ,  $r = 0 \rightarrow \infty$ .

where

$$\psi_k(z) = c_k + c_{k-1}z + \dots + c_0z^k, \quad k = 0 \rightarrow n-1 \quad (32)$$

Thus,  $f(z)$  and  $g(z)$  are constant multiples of the numerator and denominator determinants<sup>(c)</sup> in Eq. (31), respectively, and all matrix inversions are obviated.

Proof. From Eq. (21)

$$g(z) = g_n(z^n - \begin{bmatrix} 1 \\ z \\ \vdots \\ z^{n-1} \end{bmatrix}' C_{n-1}^{-1} \begin{bmatrix} c_n \\ c_{n+1} \\ \vdots \\ c_{2n-1} \end{bmatrix}) \quad (33)$$

Hence, invoking a well known result from the theory of determinants,<sup>3</sup> we find that under the normalization  $g_0 = 1$ ,

$$g_n = (-1)^{n-1} \cdot \frac{\det C_{n-1}}{\det H_n} \quad (34)$$

and  $g(z)$  equals the denominator determinant in Eq. (31) multiplied by  $(-1)^{n-1}/\det H_n$ .

From Eq. (18)

$$\begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{bmatrix} = \begin{bmatrix} c_0 & c_1 & \dots & c_{n-1} \\ & c_0 & \dots & c_{n-2} \\ & & \ddots & \vdots \\ & & & c_0 \end{bmatrix} \cdot \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix} \quad (35)$$

and from Eq. (21), with  $g_0 = 1$ ,

$$\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix} = -H_n^{-1} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} \quad (36)$$

<sup>(c)</sup>  $|A| = \det A = \text{determinant of matrix } A.$

Thus, by elimination,

$$f(z) = - \begin{bmatrix} 1 \\ z \\ \vdots \\ z^{n-1} \end{bmatrix}' \begin{bmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ & c_0 & \cdots & c_{n-2} \\ & & \ddots & \vdots \\ & & & c_0 \end{bmatrix} H_n^{-1} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix}, \quad (37)$$

which is easily recognized as the numerator determinant in Eq. (31) multiplied by  $1/\det H_n$ . Consequently,  $(-1)^{n-1}W(z)/z$  is given by Eq. (31), Q.E.D.

In many typical problems of identification the known integer  $n$  is such that a priori, degree  $H(s) \leq n$ . This case is considerably more involved in its numerical aspects than the case degree  $H(s) = n$  which falls within the scope of Theorem 1. Nevertheless, it is completely encompassed by Theorem 2.

**Theorem 2.** The degree of  $W(z)$  is less than or equal to the prescribed integer  $n (\geq 1)$ , iff the real symmetric Hankel array

$$C_n = \begin{bmatrix} c_0 & c_1 & \cdots & c_n \\ c_1 & c_2 & \cdots & c_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_n & c_{n+1} & \cdots & c_{2n} \end{bmatrix} \quad (38)$$

possesses the following properties:

- 1) For some  $p$ ,  $1 \leq p \leq n$ ,

$$\text{rank} \begin{bmatrix} c_0 & c_1 & \cdots & c_{p-1} \\ c_1 & c_2 & \cdots & c_p \\ \vdots & \vdots & \ddots & \vdots \\ c_{p-1} & c_p & \cdots & c_{2p-2} \end{bmatrix} = \text{rank} \begin{bmatrix} c_1 & c_2 & \cdots & c_p \\ c_2 & c_3 & \cdots & c_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_p & c_{p+1} & \cdots & c_{2p-1} \end{bmatrix} = p; \quad (39)$$

i.e.,

$$\text{rank } C_{p-1} = \text{rank } H_p = p. \quad (40)$$

- 2) Every principal submatrix obtained by bordering  $C_{p-1}$  with one row and the same one column and two rows and the same two columns is singular. <sup>(d)</sup>

<sup>(d)</sup> Observe that  $C_{p-1}$  is the principal  $p \times p$  submatrix of  $C_n$  located in the upper left-hand corner.

Proof. The necessity of 1) and 2) is a consequence of Theorem 1 and footnote (b). As for their sufficiency, we first invoke a result concerning the rank of a symmetric matrix<sup>4</sup> to conclude that  $\text{rank } C_n = p$ .

According to Theorem 1, the rank equalities in Eq. (39) guarantee the existence of a solution  $\underline{\zeta} = (1, g_1, \dots, g_p)'$ ,  $g_p \neq 0$ , of the homogeneous system

$$C_p \underline{\zeta} = 0_{p+1} \quad (41)$$

By substituting this  $\underline{\zeta}$  into Eq. (18) with  $n = p$ , we generate a rational function

$$W(z) = \frac{z(f_0 + f_1 z + \dots + f_{p-1} z^{p-1})}{1 + g_1 z + \dots + g_p z^p} \quad (42)$$

of degree  $p$  whose Laurent expansion automatically duplicates the given samples  $c_0, c_1, \dots, c_{2p}$ . It must now be shown that  $c_{2p+1}, c_{2p+2}, \dots, c_{2n}$  are also reproduced correctly if  $p$  is less than  $n$ .

For this purpose let us write  $C_{p+1}$  in the partitioned form

$$C_{p+1} = \left[ \begin{array}{cc|cc} & & c_p & c_{p+1} \\ & & \vdots & \vdots \\ & C_{p-1} & c_{2p-1} & c_{2p} \\ \hline c_p & \dots & c_{2p-1} & c_{2p} \\ c_{p+1} & \dots & c_{2p} & c_{2p+1} \\ & & c_{2p+1} & c_{2p+2} \end{array} \right] \quad (43)$$

Clearly,  $\text{rank } C_n = p$  implies that  $\text{rank } C_{p+1} = p = \text{rank } C_{p-1}$  and invoking another matrix result,<sup>(e)</sup>

$$\begin{bmatrix} c_{2p} & c_{2p+1} \\ c_{2p+1} & c_{2p+2} \end{bmatrix} = \begin{bmatrix} c_p & \dots & c_{2p-1} \\ c_{p+1} & \dots & c_{2p} \end{bmatrix} \cdot C_{p-1}^{-1} \cdot \begin{bmatrix} c_p & c_{p+1} \\ \vdots & \vdots \\ c_{2p-1} & c_{2p} \end{bmatrix} \quad (44)$$

<sup>(e)</sup> Let the  $r \times r$  matrix  $A$  be nonsingular. Then

$$\text{rank} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = r$$

iff  $D = CA^{-1}B$  (see Reference 3).

This means that  $c_{2p+1}$  and  $c_{2p+2}$  are uniquely determined by  $c_0, c_1, \dots, c_{2p}$  from the requirement  $\text{rank } \hat{C}_n = p$ . However, the Hankel matrix  $\hat{C}_n$  generated by the Laurent expansion of the rational function Eq. (42) has  $C_p$  in its upper left-hand corner and also satisfies  $\text{rank } \hat{C}_n = p$ .<sup>(b)</sup> Thus,  $\hat{c}_{2p+1} = c_{2p+1}$  and  $\hat{c}_{2p+2} = c_{2p+2}$ . Repeating this process on  $C_{p+2}$  it is seen that  $\hat{c}_{2p+3} = c_{2p+3}$  and  $\hat{c}_{2p+4} = c_{2p+4}$ , etc., until  $\hat{c}_{2n} = c_{2n}$ , Q.E.D.

By making contact with the modern concept of degree in terms of Hankel matrices, Theorems 1 and 2 succeed in giving Prony's algorithm<sup>5</sup> the proper system orientation. The advantages of such a setting are of course obvious. Our next theorem contains an effective explicit solution for the A coefficients in the partial fraction expansion Eq. (16b) as functions of the B's in Equation (16a). It therefore brings to a close the entire circle of ideas originated in Theorems 1 and 2.

**Theorem 3.** For any two nonnegative integers  $k$  and  $l$ , let<sup>(f)</sup>

$$S_{k,l} = \sum_{r=0}^k (-1)^r {}^k C_r r^l; \quad S_{0,0} \equiv 1 \quad (45)$$

Then, for every  $i = 1 \rightarrow q$ ,

$$A_1^{(i)} = B_1^{(i)} \quad (46)$$

$$A_{m_i}^{(i)} = B_{m_i}^{(i)} / (z_i T)^{m_i-1} \quad (47)$$

and

$$A_k^{(i)} = \frac{1}{(z_i T)^{k-1}} \cdot \frac{(-1)^{\frac{k(k-1)}{2}} \cdot (-1)^{\frac{m_i(m_i-1)}{2}}}{k! (k+1)! \dots (m_i-1)!} \cdot \begin{vmatrix} B_k^{(i)} & S_{k-1,k} & S_{k-1,k+1} & \dots & S_{k-1,m_i-1} \\ \frac{B_{k+1}^{(i)}}{(-z_i)} & S_{k,k} & S_{k,k+1} & \dots & S_{k,m_i-1} \\ \frac{B_{k+2}^{(i)}}{(-z_i)^2} & 0 & S_{k+1,k+1} & \dots & S_{k+1,m_i-1} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{B_{m_i}^{(i)}}{(-z_i)^{m_i-k}} & 0 & 0 & \dots & S_{m_i-1,m_i-1} \end{vmatrix} \quad (48)$$

$k = 2 \rightarrow m_i - 1$

<sup>(f)</sup> The binomial coefficient  ${}^k C_r$  denotes the number of combinations of  $k$  things taken  $r$  at a time and the sums  $S_{k,l}$  are discussed in Section E.

For example, if  $m_i = 1$  ( $z_i$  is simple),

$$A_1^{(i)} = B_1^{(i)} \quad ; \quad (49)$$

If  $m_i = 2$  ( $z_i$  is double),

$$A_1^{(i)} = B_1^{(i)} \quad , \quad (50)$$

$$A_2^{(i)} = B_2^{(i)} / z_i T \quad (51)$$

and if  $m_i = 3$  ( $z_i$  is triple), (g)

$$A_1^{(i)} = B_1^{(i)} \quad , \quad (52)$$

$$A_2^{(i)} = \frac{1}{2z_i T} \cdot \begin{vmatrix} B_2^{(i)} & S_{1,2} \\ B_3^{(i)} & S_{2,2} \end{vmatrix} = \frac{1}{z_i T} \left( B_2^{(i)} - \frac{B_3^{(i)}}{2z_i} \right) \quad , \quad (53)$$

$$A_3^{(i)} = B_3^{(i)} / (z_i T)^2 \quad . \quad (54)$$

Proof. Let

$$P_i(t) = \sum_{k=0}^{m_i-1} P_k^{(i)} t^k \quad , \quad i = 1 \rightarrow q \quad . \quad (55)$$

By comparing Eqs. (4) and (16b) it is easily seen that

$$A_{k+1}^{(i)} = P_k^{(i)} \cdot k! \quad , \quad k = 0 \rightarrow m_i - 1 ; \quad i = 1 \rightarrow q \quad . \quad (56)$$

Moreover, it follows from Eqs. (11) and (16a) that

$$\left( P_i \left( \frac{\partial}{\partial s} \right) \frac{1}{z - e^{sT}} \right)_{s=s_i} = \sum_{k=1}^{m_i} \frac{B_k^{(i)}}{(z - z_i)^k} \quad , \quad i = 1 \rightarrow q \quad , \quad (57)$$

and the obvious first step is to find the form of

(g) From Section E,  $S_{0,0} = 1$ ,  $S_{0,\ell} = 0$ ,  $\ell > 0$ ,  $S_{1,2} = -1$ ,  $S_{2,2} = 2$  and  $S_{k,k} = (-1)^k k!$ ,  $k = 0 \rightarrow \infty$ .

$$I_{\ell}^{(i)} \equiv \left( \frac{\partial^{\ell}}{\partial s^{\ell}} \frac{1}{z - e^{sT}} \right)_{s=s_i} \quad (58)$$

for any nonnegative integer  $\ell$ .

Let  $y = e^{sT}$ . Clearly,

$$\frac{\partial}{\partial s} = T e^{sT} \frac{\partial}{\partial y} = T y \frac{\partial}{\partial y} \quad (59)$$

and therefore, operationally,

$$\frac{\partial^{\ell}}{\partial s^{\ell}} \equiv \left( \frac{\partial}{\partial s} \right)^{\ell} = T^{\ell} \left( y \frac{\partial}{\partial y} \right)^{\ell} \quad (60)$$

Hence, applying the well-known<sup>6</sup> Liebnitz-type rule

$$\left( y \frac{\partial}{\partial y} \right)^{\ell} = \sum_{k=0}^{\ell} \frac{(-1)^k}{k!} S_{k,\ell} y^k \left( \frac{\partial}{\partial y} \right)^k \quad (61)$$

to the function  $1/(z-y)$  we obtain, in view of Eq. (58),

$$I_{\ell}^{(i)} = \sum_{k=0}^{\ell} T^{\ell} (-1)^k z_i^k S_{k,\ell} (z - z_i)^{-(k+1)} \quad (62)$$

Now, by taking Eq. (56) into account and grouping terms in Eq. (57) it is readily verified that

$$B_{k+1}^{(i)} = \sum_{\ell=k}^{m_i-1} P_{\ell} T^{\ell} (-1)^k z_i^k S_{k,\ell} = (-1)^k z_i^k \cdot \sum_{\ell=k}^{m_i-1} \frac{T^{\ell}}{\ell!} S_{k,\ell} A_{\ell+1}^{(i)} \quad (63)$$

$$k = 0 \rightarrow m_i - 1$$

Rewritten in matrix form, Eq. (63) goes into the upper-triangular linear system

$$\begin{bmatrix} B_1^{(i)} \\ \frac{B_2^{(i)}}{(-z_i)} \\ \frac{B_3^{(i)}}{(-z_i)^2} \\ \vdots \\ \frac{B_{m_i}^{(i)}}{(-z_i)^{m_i-1}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & S_{1,1} & S_{1,2} & \dots & S_{1,m_i-1} \\ 0 & 0 & S_{2,2} & \dots & S_{2,m_i-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & S_{m_i-1,m_i-1} \end{bmatrix} \begin{bmatrix} A_1^{(i)} \\ \frac{T A_2^{(i)}}{1!} \\ \frac{T^2 A_3^{(i)}}{2!} \\ \vdots \\ \frac{T^{m_i-1} A_{m_i}^{(i)}}{(m_i-1)!} \end{bmatrix}, \quad (64)$$

and Eqs. (46) to (48) drop out immediately by Cramer's rule, Q.E.D.

Comment. The numerical problems that are encountered in performing partial fraction expansions are by no means trivial. However, the literature on this subject is rather abundant and we will not pursue the topic further.

### C. Worked Examples

We have chosen two examples to illustrate some of the mechanics associated with the procedures developed in the above theorems.

Example 1. Show how to find the minimum-degree transfer functions  $H(s)$  which generate the samples

$$c_0 = 1, \quad c_1 = 2, \quad c_2 = 0, \quad c_3 = 0 \quad (65)$$

Solution. Ordinarily, with either  $2n$  or  $2n+1$  samples we attempt to construct  $W(z)$  so that its degree is less than or equal to  $n$  and it is therefore appropriate to study  $C_n$ . Since  $4 = 2 \times 2$ , consider

$$C_2 = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & c_4 \end{bmatrix}, \quad (66)$$

where the sample  $c_4$  is unknown. According to Theorem 1, degree  $W(z) = 2$  iff  $c_4 = 0$  and

$$\text{rank} \begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix} = \text{rank} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} = 2, \quad (67)$$

which is impossible. Clearly, degree  $W(z) = 1$  is also precluded so that degree  $W(z) \geq 3$ . Consequently, consider

$$C_3 = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 2 & 0 & 0 & c_4 \\ 0 & 0 & c_4 & c_5 \\ 0 & c_4 & c_5 & c_6 \end{bmatrix}, \quad (68)$$

where  $c_4$ ,  $c_5$  and  $c_6$  are as yet undermined.

According to Theorem 1, degree  $W(z) = 3$  is achieved iff

$$\det C_3 = 4(c_5^2 - c_4 c_6) - c_4^3 = 0 \quad (69)$$

and

$$\det C_2 = -4c_4 \neq 0; \quad \det H_2 = -2c_4^2 \neq 0 \quad (70)$$

Thus, we may choose any  $c_4 \neq 0$  and then select  $c_5$  and  $c_6$  so that

$$c_4^3 + 4c_4 c_6 - 4c_5^2 = 0. \quad (71)$$

The particular choice  $c_4 = 4$ ,  $c_5 = 0$ ,  $c_6 = -4$  yields the equation,

$$\begin{bmatrix} 1 & 2 & 0 & 0 \\ 2 & 0 & 0 & 4 \\ 0 & 0 & 4 & 0 \\ 0 & 4 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} = \underline{0}_4 \quad (72)$$

for  $\underline{\zeta} = (1, g_1, g_2, g_3)'$  whose solution is given by

$$g_2 = 0; \quad g_1 = g_3 = -\frac{1}{2} \quad (73)$$

$$g(z) = 1 - \frac{z}{2} - \frac{z^3}{2}. \quad (74)$$

The polynomial  $g(z)$  possesses the three simple zeros,

$$z_1 = 1 ; z_{2,3} = \frac{-1 \pm j\sqrt{7}}{2} . \quad (75)$$

From Eq. (18),

$$f_2 = -\frac{1}{2} , f_1 = -1 , f_0 = -\frac{1}{2} \quad (76)$$

and

$$\frac{W(z)}{2} = \frac{\frac{z^2}{2} + z + \frac{1}{2}}{\frac{z^3}{2} + \frac{z}{2} - 1} = \frac{z^2 + 2z + 1}{z^3 + z - 2} = \frac{(z+1)^2}{(z-1)(z^2 + z + 2)} , \quad (77)$$

$$= \frac{1}{z-1} - \frac{j}{\sqrt{7}} \cdot \frac{1}{z-z_2} + \frac{j}{\sqrt{7}} \cdot \frac{1}{z-\bar{z}_2} . \quad (78)$$

Thus, with  $T=1$ ,

$$s_1 = \ln 1 = 0 ,$$

$$s_{2,3} = \ln z_{2,3} = 0.34657 \pm j 1.93216 ,$$

$$H(s) = \frac{1}{s} - \frac{j}{\sqrt{7}} \cdot \frac{1}{s-s_2} + \frac{j}{\sqrt{7}} \cdot \frac{1}{s-\bar{s}_2} \quad (79)$$

and

$$h(t) = 1 + 0.75593e^{0.34657t} \sin(1.93216)t . \quad (80)$$

A check shows that to at least five places,  $h(0) = 1$ ,  $h(1) = 2$  and  $h(2) = h(3) = 0.00001$ .

It is easily shown that the most general denominator  $g(z)$  compatible with the constraint  $c_4 \neq 0$  and the equality in Eq. (68) is given by

$$g(z) = 1 - \frac{z}{2} + \frac{c_5}{c_4} z^2 - \frac{2}{c_4} z^3 . \quad (81)$$

This means that the four samples  $c_0 = 1$ ,  $c_1 = 2$ ,  $c_2 = 0$  correspond to an infinity of different pole assignments for the minimum degree  $H(s)$  consistent with these samples.

Nevertheless, we maintain that degree  $H(s) \geq 3$  is the only reasonable conclusion that can be drawn from the given data. For if the degree of the "true"  $H(s)$  is actually two, say, the matrix

$$H_2 = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \quad (82)$$

must represent a noise-corrupted version of some

$$(H_2)_{\text{true}} = \left[ \begin{array}{c|c} 2 + \epsilon_1 & \epsilon_2 \\ \hline \epsilon_2 & \epsilon_3 \end{array} \right] \quad (83)$$

where

$$(2 + \epsilon_1)\epsilon_3 - \epsilon_2^2 \neq 0 \quad (84)$$

In other words, the continuous noise components  $n_1, n_2, n_3$  introduced in the measurement of  $h(1), h(2)$  and  $h(3)$  must satisfy the equalities  $n_1 = -\epsilon_1, n_2 = -\epsilon_2$  and  $n_3 = -\epsilon_3$ . But such a triplet constitutes an event which lies on the zero-probability surface,

$$(2 + \epsilon_1 + n_1)(\epsilon_3 + n_3) - (\epsilon_2 + n_2)^2 = 0 \quad (85)$$

and the possibility of its occurrence must be discarded.

**Example 2.** Find the minimum-degree  $H(s)$  corresponding to the eleven noise-free samples

$$c_0 = 0, c_1 = \frac{17}{16}, c_2 = \frac{1}{4}, c_3 = -\frac{7}{16}, c_4 = 1 \quad (86)$$

$$c_5 = \frac{41}{16}, c_6 = \frac{9}{4}, c_7 = \frac{33}{16}, c_8 = 4, c_9 = \frac{97}{16}, c_{10} = \frac{25}{4}$$

spaced  $T = 1/4$  secs. apart.

**Solution.** Since  $11 = 2 \times 5 + 1$ , consider

$$C_5 = \begin{bmatrix} 0 & \frac{17}{16} & \frac{1}{4} & -\frac{7}{16} & 1 & \frac{41}{16} \\ \frac{17}{16} & \frac{1}{4} & -\frac{7}{16} & 1 & \frac{41}{16} & \frac{9}{4} \\ \frac{1}{4} & -\frac{7}{16} & 1 & \frac{41}{16} & \frac{9}{4} & \frac{33}{16} \\ -\frac{7}{16} & 1 & \frac{41}{16} & \frac{9}{4} & \frac{33}{16} & 4 \\ 1 & \frac{41}{16} & \frac{9}{4} & \frac{33}{16} & 4 & \frac{97}{16} \\ \frac{41}{16} & \frac{9}{4} & \frac{33}{16} & 4 & \frac{97}{16} & \frac{25}{4} \end{bmatrix} \quad (87)$$

Because of the nonsingularity of  $C_4$ , degree  $H(s) \geq 5$  (Theorem 1) and it is easily checked<sup>(h)</sup> that

$$C_5 \underline{\zeta} = \underline{0}_6 \quad (88)$$

where  $\underline{\zeta} = (1, -3, 4, -4, 3, -1)'$ . Hence,  $\det C_5 = 0$ . In addition, direct computation shows that  $\det H_5 \neq 0$  and from Theorem 1, degree  $H(s) = \text{degree } W(z) = 5$ .

The polynomial  $g(z)$  is given by

$$g(z) = 1 - 3z + 4z^2 - 4z^3 + 3z^4 - z^5 = (1 - z)^3 (z^2 + 1) \quad (89)$$

and  $H(s)$  possesses  $s_1 = 0$  as a pole of multiplicity three and the two simple complex conjugate poles,  $s_{2,3} = \pm 4 \ln j = \pm 2\pi j$ . Hence,  $q = 3$ ,  $m_1 = 3$  and  $m_2 = m_3 = 1$ .

From the recursions in Eq. (18),

$$f(z) = \frac{1}{16} (15 - 49z + 47z^2 - 17z^3) \quad (90)$$

whence (all details are omitted),

$$\frac{W(z)}{z} = \frac{15 - 49z + 47z^2 - 17z^3}{16(1 - z)^3 (z^2 + 1)} \quad (91)$$

$$= \frac{j/2}{z+j} - \frac{j/2}{z-j} + \frac{1}{16} \cdot \frac{1}{(z-1)^2} + \frac{1}{8} \cdot \frac{1}{(z-1)^3} \quad (92)$$

Thus,

$$H(s) = \frac{j/2}{s+2\pi j} - \frac{j/2}{s-2\pi j} + \frac{A_1^{(1)}}{s} + \frac{A_2^{(1)}}{s^2} + \frac{A_3^{(1)}}{s^3} \quad (93)$$

$$= \frac{2\pi}{s^2 + (2\pi)^2} + \frac{A_1^{(1)}}{s} + \frac{A_2^{(1)}}{s^2} + \frac{A_3^{(1)}}{s^3} \quad (94)$$

By straight substitution of  $B_1^{(1)} = 0$ ,  $B_2^{(2)} = 1/16$  and  $B_3^{(1)} = 1/8$  into Eqs. (52) to (54) we get

<sup>(h)</sup>The data  $c_0, c_1, \dots, c_{10}$  has been generated by sampling

$$h(t) = t^2 + \sin 2\pi t \quad .$$

$$A_1^{(1)} = 0 \quad ,$$

$$A_2^{(1)} = 4\left(\frac{1}{16} - \frac{1}{16}\right) = 0 \quad ,$$

$$A_3^{(1)} = 16B_3^{(1)} = 2 \quad .$$

$$\therefore H(s) = \frac{2}{s^3} + \frac{2\pi}{s^2 + (2\pi)^2} \quad (95)$$

and

$$h(t) = t^2 + \sin 2\pi t \quad , \quad (96)$$

which agrees with footnote (h).

#### D. The Effect of Measurement Noise

In essence, the theoretical problem of identification is that solving an equation

$$A(x) = y \quad (97)$$

for  $x$ , given  $y$  and the operator  $A$ . If the inverse  $A^{-1}$  exists, the solution is given by the rule

$$x = A^{-1}(y) \quad (98)$$

which is applicable to every  $y$  in the range of  $A$ . However, due to the inevitable presence of measurement noise, Eq. (98) may not constitute a practical solution.

Indeed, suppose that  $y_0$  is the "true" image of  $x_0$  and that because of noise disturbance  $d$ ,  $y_0$  is corrupted into  $y = y_0 + d$ . Then, any scheme used to implement Eq. (98) will return (in exact arithmetic),

$$x = A^{-1}(y_0 + d) \quad (99)$$

instead of

$$x_0 = A^{-1}(y_0) \quad (100)$$

and it is therefore important to assess the magnitude of the error

$$\epsilon = \|x - x_0\| \quad (101)$$

in some appropriate norm,  $\|, \|$ . In particular, it is clear that a stable inversion procedure must be such that  $\epsilon$  tends to zero as  $\|d\| \rightarrow 0$ . Or, stated differently, the

identification problem is well-posed if the inverse operator  $A^{-1}$  is continuous.

**Theorem 4.** Let the total number of poles  $n$  of a transfer function  $H(s)$  of type (1) be known in advance. Then, if  $h(t) \Leftrightarrow H(s)$ , the problem of restoring  $h(t)$  from  $2n+1$  of its samples  $c_k = h(kT)$ ,  $k = 0 \rightarrow 2n$ , is well-posed if the sampling rate  $1/T$  is high enough to preclude folding.

**Proof.** According to the corollary to Theorem 1, in the absence of noise the number of distinct  $z_i$ 's,  $q$  say, and their respective multiplicities  $m_i$ ,  $i = 1 \rightarrow q$ , are determined as the zeros of the denominator determinant in Equation (31). From the  $z_i$ 's we obtain the  $q$  distinct exponents  $s_i = \frac{1}{T} \ln z_i$ ,  $i = 1 \rightarrow q$ , and it follows that the only unknowns left in  $h(t)$  are the  $m_1 + m_2 + \dots + m_q = n$  coefficients of  $P_1(t)$ ,  $P_2(t)$ ,  $\dots$ ,  $P_q(t)$ .

In turn, the vector of these  $n$  coefficients is determined as the solution of the  $n \times n$  linear system,

$$c_k = \sum_{i=1}^q P_i(kT) e^{s_i kT}, \quad k = 0 \rightarrow n-1. \quad (102)$$

Moreover, this solution is unique because the system determinant is a generalized Vandermonde which cannot vanish.

Let us continue to use the zeros of the denominator determinant in Eq. (31) as our estimates for the  $z_i$ 's even in the presence of noise. By continuity, as the noise becomes evanescent, these roots coalesce into distinct nonoverlapping clusters. The number of such clusters yields  $q$  and a count of the number of  $z_k$ 's in the  $i^{\text{th}}$  cluster yields  $m_i$ ,  $i = 1 \rightarrow q$ .

As an approximation to the noise-free  $z_i$  we choose the center of gravity  $\zeta_i$  of all the roots in the  $i^{\text{th}}$  cluster. Clearly, as the noise goes to zero,  $\zeta_i \rightarrow (z_i)_{\text{true}}$ ,  $i = 1 \rightarrow q$ , and  $c_k \rightarrow (c_k)_{\text{true}}$ ,  $k = 0 \rightarrow n-1$ . Consequently, the solution of Eq. (102) tends to the true coefficients and the impulse response calculated by means of this algorithm has the true  $h(t)$  as its limit, Q.E.D.

**Comment 1.** The technique employed in the proof of Theorem 4 bypasses the need to determine  $f(z)$  and the partial fraction expansion of  $f(z)/g(z)$ . Note however, that in the solution of Eq. (102) only  $n$  of the  $2n+1$  samples are used, and the question of which set of  $n$  to choose is left unanswered. At this point it becomes necessary to take the statistical properties of the noise into account.

**Comment 2.** Unlike several other methods, Prony's approach has the great merit of exploiting the finite-dimensional nature of the problem from the outset and it is this

that underlies the validity of Theorem 4. Nevertheless, the algorithm can exhibit severe ill-conditioning, especially in the presence of multiple poles. The availability of the explicit formulas (45) to (48) derived in Theorem 3 should help relieve the situation.

Comment 3. The original work by Prony was published in 1795, the same year Gauss presented his theory of least squares estimation. As we have seen, the technique is ideally suited for modal analysis and in the hands of Van Blaricum and his co-workers,<sup>7,8</sup> has proven itself to be an effective diagnostic tool for exploring the complex resonances of a system. It continues to stimulate researchers in many different fields.

#### E. Appendix

For any two nonnegative integers  $k$  and  $\ell$  let  $a_{k,\ell}$  denote the coefficient of  $x^\ell$  in the power series expansion of

$$a(x) = (1 - e^x)^k \quad (103)$$

about  $x=0$ . By the Binomial theorem,

$$a(x) = \sum_{r=0}^k (-1)^r C_r^k e^{rx}$$

whence,

$$a_{k,\ell} = \frac{1}{\ell!} \cdot \sum_{r=0}^k (-1)^r C_r^k r^\ell = \frac{1}{\ell!} S_{k,\ell} \quad (104)$$

However, since

$$e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!} \quad (105)$$

Equation (103) is also expressible as

$$a(x) = (-x)^k \left(1 + \frac{x}{2!} + \frac{x^2}{3!} + \dots\right)^k \quad (106)$$

Thus,

$$a_{k,\ell} = S_{k,\ell} = 0 \quad , \quad \ell < k \quad (107)$$

and

$$S_{k,\ell} = (-1)^k \ell! b_{k,\ell-k}, \quad \ell \geq k, \quad (108)$$

where  $b_{k,\ell}$  is the coefficient of  $x^\ell$  in the expansion of

$$\left(1 + \frac{x}{2!} + \frac{x^2}{3!} + \dots\right)^k. \quad (109)$$

But

$$\begin{aligned} \left(1 + \frac{x}{2!} + \frac{x^2}{3!} + \dots\right)^k &= 1 + \frac{k}{2} x + \frac{k(3k+1)}{4!} x^2 + \frac{k^2(k+1)}{2 \cdot 4!} x^3 + \\ &+ \frac{k}{10(4!)^2} (15k^3 + 30k^2 + 5k - 2)x^4 + \\ &+ \frac{k^2}{20(4!)^2} (3k^3 + 10k^2 + 5k - 2)x^5 + \\ &+ \frac{k}{7!(4!)^2} (63k^5 + 315k^4 + 315k^3 - 91k^2 - 42k + 16)x^6 + \\ &+ \dots \end{aligned} \quad (110)$$

For possible future reference we can therefore generate the list  $S_{k,k+r}$ ,  $r = 0 \rightarrow 8$ :

$$S_{k,k} = (-1)^k k! \quad (111)$$

$$S_{k,k+1} = (-1)^k (k+1)! \frac{k}{2}, \quad (112)$$

$$S_{k,k+2} = (-1)^k (k+2)! \frac{k}{4!} (3k+1), \quad (113)$$

$$S_{k,k+3} = (-1)^k (k+3)! \frac{k^2}{2 \cdot 4!} (k+1), \quad (114)$$

$$S_{k,k+4} = (-1)^k (k+4)! \frac{k}{10(4!)^2} (15k^3 + 30k^2 + 5k - 2), \quad (115)$$

$$S_{k,k+5} = (-1)^k (k+5)! \frac{k^2}{20(4!)^2} (3k^3 + 10k^2 + 5k - 2), \quad (116)$$

$$S_{k,k+6} = (-1)^k (k+6)! \frac{k}{7!(4!)^2} (63k^5 + 315k^4 + 315k^3 - 91k^2 - 42k + 16), \quad (117)$$

$$S_{k,k+7} = (-1)^k (k+7)! \frac{k^2}{3!4!8!} (9k^6 + 65k^5 + 105k^4 - 7k^3 - 4074k^2 + 12112k - 8064) , \quad (118)$$

$$S_{k,k+8} = (-1)^k (k+8)! \frac{k}{5!9!(2!)} (135k^7 + 1260k^6 + 3150k^5 + 840k^4 - 2345k^3 + 540k^2 + 404k - 144) . \quad (119)$$

With the aid of formulas (111) to (119) it is a straightforward matter to program Eq. (48) for multiplicities  $m_i \leq 9$ .

SCEEE

Post-Doctoral Program

D. C. Youla

#### REFERENCES

1. "Theory of Servomechanisms," MIT Radiation Lab. Series, Vol. 25, McGraw-Hill, New York (1947).
2. D. C. Youla and Plinio Tissi, "n-Port Synthesis via Reactance Extrancion," Part I; 1966 IEEE International Convention Record, Part 7.
3. F. R. Gantmacher, "The Theory of Matrices," Vol. I, Chelsea Publishing Co., New York (1960).
4. Maxime Bocher, "Introduction to Higher Algebra," The McMillan Co., (1907).
5. F. B. Hildebrand, "Introduction to Numerical Analysis," McGraw-Hill, New York (1956).
6. I. J. Schwatt, "Operations with Infinite Series," Chelsea Publishing Co., New York (1924).
7. Michael Lee Van Blaricum, "Techniques for Extracting the Complex Resonances of a System Directly from its Transient Response," Ph.D. Dissertation, Dept. of Electrical Engineering, University of Illinois, Dec. 1975; also AFWL Inter-action Note No. 301. (Contains a good bibliography.)
8. M. L. Van Blaricum and R. Mittra, "Problems and Solutions Associated with Prony's Method for Processing Transient Data," IEEE Trans. on Ant. and Prop., pp. 174-182 (Jan. 1978).

## ON ORDER DETERMINATION OF LINEAR AR MODELS

F. Nakajima and F. Kozin

A. Introduction

We are often faced with the problem of estimating parameters and the order of a model that will fit a given set of observations. Akaike<sup>1</sup> established a procedure, which estimates order as well as unknown parameters, by introducing the mean log-likelihood as a measure of fit and by maximizing this measure of fit. However, Akaike's formulation requires the knowledge of the probability law of the observations. Following our approach to characterization of consistent estimators,<sup>2</sup> in this report, we shall define a criterion as a measure of fit which can be formulated without knowledge of the probability law of observations and which has a close relation with the formulation for the parameter estimation problem established in our companion report on "Characterization of Consistent Estimates." Using this criterion, we shall obtain a procedure to estimate the order and the parameters present in linear time-varying AR models.

B. A Criterion for Order Determination

We assume that the given data  $\{y_t\}$  are generated by the following linear AR model:

$$y_t + \overset{\circ}{a}_1(t-1)y_{t-1} + \dots + \overset{\circ}{a}_k(t-k)y_{t-k} = b_t w_t \quad (1)$$

$$y_{-1} = y_{-2} = \dots = -y_{-k}^{\circ} = 0$$

where  $\{w_t\}$  is an  $n \times 1$  independent sequence with zero mean and variance  $I$  and  $\{y_t\}$  is the  $n \times 1$  output. In the system Eq. (1),  $\{\overset{\circ}{a}_i(t), i = 1, 2, \dots, k\}$  and  $k$  represent unknown coefficients and unknown system order, respectively.

We shall represent the time-varying coefficient  $\{\overset{\circ}{a}_i(t)\}$  as

$$\overset{\circ}{a}_i(t) = \sum_{j=1}^m \overset{\circ}{a}_{ij} F_{ij}(t)$$

for each  $i = 1, 2, \dots, k$  where  $\{\overset{\circ}{a}_{ij}\}$  are unknown constant parameters and  $\{F_{ij}(t)\}$  are known  $n \times n$  uniformly bounded matrices. Furthermore, we assume that  $k \in \{1, 2, \dots, \ell\}$ , where  $\ell$  is a possible largest order, the true unknown parameter  $\overset{\circ}{\theta}_L = (\overset{\circ}{a}_{11}, \dots, \overset{\circ}{a}_{1m}, \overset{\circ}{a}_{21}, \dots, \overset{\circ}{a}_{km}, 0, \dots, 0)^T$  belongs to a compact set  $\Theta$  in  $R^L$ ,  $L = \ell \times m$ ,  $\Theta_L = \{\theta_L \mid \|\theta_L\| \leq c, c < \infty\}$  and  $\theta_L = (\theta_{L1}, \theta_{L2}, \dots, \theta_{L,L})^T = (a_{11}, a_{12}, \dots, a_{\ell m})^T$ .

The problem of model fitting is often formulated as one of the selection of the probability density function (p.d.f.),  $P(Y_N | \theta_K)$  based on observations  $Y_N$  given  $\theta_K$  where  $Y_N = \{y_t, t = 0, 1, \dots, N-1\}$  and  $\theta_K$  is a parameter restricted to the space,  $\theta_K = (\theta_{K,1}, \dots, \theta_{K,K}, 0, \dots, 0)^T$  where  $K = k \times l$ ,  $1 \leq k \leq l$ . We call  $k$  the order of the model. To discriminate between two systems, where one is the true system and the other is the estimated model, Akaike introduced the Kullback-Leibler information criterion (KLC) for each order  $k$  defined by  $KLC(\hat{\theta}_L; \theta_K; N) = E\{\log P(Y_N | \hat{\theta}_L) - \log P(Y_N | \theta_K)\}$ . Since KLC takes the value zero only at  $\theta_K = \hat{\theta}_L$  and otherwise positive, then the best estimate of the order can be obtained by minimizing KLC with respect to the order  $k \in \{1, 2, \dots, l\}$ .

In this report we shall solve the problem of order determination without knowledge of the p.d.f. of  $Y_N$ . To do this, we choose a functional  $q_N(\theta_K)$  of  $Y_N$ ,  $\theta_K$  and  $N$  which satisfies the following conditions

$$(a-1) \quad \lim_{N \rightarrow \infty} E\{q_N(\theta_K)\} \text{ exists.}$$

$$(a-2) \quad \lim_{N \rightarrow \infty} [q_N(\theta_K) - E\{q_N(\theta_K)\}] = 0 \text{ w.p.1.}$$

$$(a-3) \quad \lim_{N \rightarrow \infty} E\{q_N(\theta_K) - q_N(\hat{\theta}_L)\} \geq 0.$$

$$\text{with equality iff } \theta_K = \hat{\theta}_L.$$

We define a criterion for discriminating two systems as

$$\begin{aligned} I(\hat{\theta}_L; \theta_K) &\triangleq \lim_{N \rightarrow \infty} E\{q_N(\theta_K) - q_N(\hat{\theta}_L)\} \\ &= \lim_{N \rightarrow \infty} (q_N(\theta_K) - q_N(\hat{\theta}_L)) \text{ w.p.1} \end{aligned} \quad (2)$$

for each order  $k$ . The above criterion  $I(\hat{\theta}_L; \theta_K)$  can be evaluated without knowledge of the p.d.f. of  $Y_N$ .

We define a compact set  $\Theta_K$  in  $R^L$  as  $\Theta_K \triangleq \{\theta_K | \|\theta_K\| \leq c, c < \infty\}$  for each order  $k$  and let us define a parameter  $\bar{\theta}_K$  in  $\Theta_K$  as

$$\min_{\theta_K \in \Theta_K} \left[ \lim_{N \rightarrow \infty} E\{q_N(\theta_K)\} \right] = \lim_{N \rightarrow \infty} E\{q_N(\bar{\theta}_K)\}.$$

Since this definition of  $\bar{\theta}_K$  implies that for each order  $k$ ,  $I(\hat{\theta}_L^0 : \theta_K) \geq I(\hat{\theta}_L^0 : \bar{\theta}_K)$ ; then  $I(\hat{\theta}_L^0 : \bar{\theta}_K)$  indicates the best fitting measure of  $Q(\theta_K)$  for  $Q(\hat{\theta}_L^0)$  where  $Q(\theta) = \lim_{N \rightarrow \infty} E\{q_N(\theta)\}$ . This fact tells us that if we can evaluate  $I(\hat{\theta}_L^0 : \bar{\theta}_K)$  for each order  $k$ , then we can obtain the best order  $\bar{k}$  such that

$$\min_{1 \leq k \leq l} I(\hat{\theta}_L^0 : \bar{\theta}_K) = I(\hat{\theta}_L^0 : \bar{\theta}_{\bar{K}}), \text{ where } \bar{K} = \bar{k} \times m.$$

Our final aim is to obtain the estimate of  $I(\hat{\theta}_L^0 : \bar{\theta}_K)$  for sufficiently large samples. In order to do this, we first show the a.s. consistency of the estimate of  $\bar{\theta}_K$  for each order  $k$ .

Assume that the system (1) satisfies

$$(b-1) \quad \hat{\theta}_L^0 \in \Theta_L \text{ and } k \in \{1, 2, \dots, l\}.$$

$$(b-2) \quad b_t \text{ is uniformly bounded and non-singular for all } t.$$

$$(b-3) \quad \text{The system (1) is exponentially stable.}$$

$$(b-4) \quad E\|w_t\|^8 \leq c_1 \text{ for all } t \geq 0.$$

$$(b-5) \quad b_t \text{ is a periodic function with period } T_0, \text{ and for each } i = 1, \dots, l \\ \{F_{ij}(t), j = 1, 2, \dots, m\} \text{ are linearly independent and are periodic functions with period } T_0.$$

We note that the periodicity condition (b-5) may be relaxed to other classes of functions for which  $\lim_{N \rightarrow \infty} E\{q_N(\theta_K)\}$  exists.

For the system Eq. (1) with conditions (b-1) to (b-5) it suffices to choose a quadratic functional  $q_N(\theta_K)$  defined as

$$q_N(\theta_K) = \frac{1}{N} \sum_{t=0}^{N-1} (\hat{y}_t^{(k)})^T (b_t b_t^T)^{-1} \hat{y}_t^{(k)} \quad (3)$$

where  $\hat{y}_t^{(k)} = y_t + a_1(t-1)y_{t-1} + \dots + a_k(t-k)y_{t-k}$  and  $\theta_K \in \Theta_K$ .

Since the functional  $q_N(\theta_K)$  satisfies the sufficient condition of a.s. consistent estimate, then we have that  $\hat{\theta}_K^N \rightarrow \bar{\theta}_K$  w.p.1 as  $N \rightarrow \infty$  where  $\hat{\theta}_K^N$  is defined by

$$\min_{\theta_K \in \Theta_K} q_N(\theta_K) = q_N(\hat{\theta}_K^N).$$

Note that  $q_N(\theta_K)$  satisfies the conditions (a-1) to (a-3) given above. Applying a central limit theorem for martingales by Brown and Eagleson, we have

$$(\hat{\theta}_K^N - \bar{\theta}_K) \stackrel{d}{\sim} N(0, \frac{1}{N} C_{K,N})$$

where  $N(0, \frac{1}{N} C_{K,N})$  denotes the normal distribution with zero mean and variance

$\frac{1}{N} C_{K,N}$  and  $C_{K,N}$  is given by

$$C_{K,N} = \frac{\partial}{\partial \theta_K} \left( \frac{\partial}{\partial \theta_K} \frac{1}{2} q_N(\theta_K) \right)^T.$$

From the above two results, we take  $\bar{W}_K^N$  defined below as an estimate of  $I(\bar{\theta}_L^0; \bar{\theta}_K)$  for sufficiently large  $N$ ,

$$\bar{W}_K^N = 2(\hat{\theta}_K^N - \bar{\theta}_K)^T C_{K,N}(\hat{\theta}_K^N - \bar{\theta}_K) + q_N(\hat{\theta}_K^N) - q_N(\bar{\theta}_L^0)$$

for each order  $k$ . Hence

$$E\{\bar{W}_K^N\} \cong -q_N(\bar{\theta}_L^0) + q_N(\hat{\theta}_K^N) + \frac{2K}{N} \quad (4)$$

where the expectation is taken with respect to the asymptotic distribution of  $\hat{\theta}_K^N$ .

The estimate of the order  $\hat{k}$  is obtained from Eq. (4) as the value of  $k$ , that minimizes

$$q_N(\hat{\theta}_K^N) + \frac{2K}{N} \quad (5)$$

National Science Foundation  
ATA-7303217

F. Nakajima and F. Kozin

#### REFERENCES

1. H. Akaike, "A New Look at the Statistical Model Identification," IEEE, AC-19, No. 6, (Dec. 1974).
2. F. Nakajima and F. Kozin, "A Characterization of Consistent Estimators," submitted to IEEE Trans. on Auto. Contr.

## CHARACTERIZATION OF CONSISTENT ESTIMATES

F. Nakajima and F. Kozin

A. Introduction

The strong consistency of parameter estimators has been studied by many researchers in recent years. Strong consistency results have been established for Maximum Likelihood Estimates (MLE),<sup>1,2</sup> Least Squares Estimates (LSE),<sup>3</sup> and more recently for Prediction Error Estimates (PEE).<sup>4,5</sup> The basic characteristic of each of these estimates is that they are defined in terms of an extremum value of some appropriate function of the observed data and the unknown parameters. The strong consistency results that are presently available, require conditions on the appropriate functions that define MLE, LSE and PEE, respectively. Conditions such as differentiability, with respect to the unknown parameters, existence of certain limits, etc., are usually required.

In this report we will present a reasonably general characterization of strong consistency which apparently allows us to treat a broader class of estimation problems than has been treated before.

We establish that strong consistency is basically a question of limits of the extremal points of a suitable sequence of functions of the observations. This sequence of functions must satisfy certain almost sure asymptotic properties, otherwise they are quite arbitrary.<sup>6</sup>

In this brief report we will simply state the main theorem and some results.

B. Characterization of Consistent Estimates

Let  $Y_N = \{y_k, k = 0, 1, \dots, N-1\}$  be a collection of observed data  $y_k$  generated by some system which is parameterized by an unknown constant parameter  $\theta$ , where  $\theta$  lies in a finite dimensional space  $R^n$ . We let  $P(Y_N/\theta)$  be the probability measure of  $Y_N$  given a true parameter  $\theta$ . We assume that  $\theta$  belongs to a compact set  $\Theta$  in  $R^n$ .

Then we have the following theorem which establishes the convergence properties of the estimates of  $\theta$ .

Theorem 1.

If there exists a functional of  $Y_N$ ,  $\theta \in \Theta$  and  $N$ , say  $q_N(\theta)$ , which satisfies the following conditions:

- (1)  $q_N(\theta)$  is measurable with respect to  $P(Y_N/\theta)$

- (2)  $\inf_{k \geq N} (q_k(\theta) - q_k(\hat{\theta}))$  converges uniformly on  $\Theta$  as  $N \rightarrow \infty$ , w.p.1
- (3)  $\lim_{N \rightarrow \infty} \inf_{k \geq N} (q_k(\theta) - q_k(\hat{\theta}))$  continuous w.p.1 on  $\Theta$
- (4)  $\lim_{N \rightarrow \infty} \inf_{k \geq N} (q_k(\theta) - q_k(\hat{\theta})) \geq 0$ , w.p.1

with equality holding iff  $\theta \in G$ , where  $G$  is some closed subset of  $\Theta$ , then,  $\hat{\theta}^N$ , converges to  $G$  w.p.1 as  $N$  tends infinity, where  $\hat{\theta}^N$  is defined by

$$\min_{\theta \in \Theta} q_N(\theta) = q_N(\hat{\theta}^N)$$

We note that if  $G$  contains only the single point  $\hat{\theta}$ , then  $\hat{\theta}^N \rightarrow \hat{\theta}$  with probability one, and, hence, the estimate is strongly consistent. We note that Theorem 1 holds in the continuous time observations,  $Y_T = \{y_t; 0 \leq t \leq T\}$ , the limits being taken as  $T$  approaches infinity. Furthermore, under suitable conditions Theorem 1 also holds in the mean.

If it can be established that the limit function

$$Q(\theta) = \lim_{N \rightarrow \infty} q_N(\theta)$$

exist with probability one, then conditions (1) to (3) of the theorem are simply stated as,

- (1)  $q_N(\theta)$  is continuous on  $\Theta$  for each  $N$ .
- (2)  $\lim_{N \rightarrow \infty} q_N(\theta) = Q(\theta)$  uniformly on  $\Theta$ .
- (3)  $Q(\theta)$  takes its minimum value on the set  $G \subset \Theta$ .

### C. Examples

As one of the possible applications of the main theorem, we consider the system,

$$x_{k+1} = \overset{\circ}{A}_k x_k + \overset{\circ}{B}_{k+1} w_{k+1}, \quad (1)$$

$$y_k = x_k + v_k$$

with initial state  $x_0$ , where  $x_k$ ,  $w_k$ ,  $y_k$  and  $v_k$  are  $n \times 1$ ,  $p \times 1$ ,  $n \times 1$  and  $n \times 1$  vectors, and  $\overset{\circ}{A}_k$ ,  $\overset{\circ}{B}_k$  are time varying  $n \times n$ ,  $n \times p$  matrices which include unknown parameters.

Assume that  $\{w_k\}$ ,  $\{v_k\}$  and  $x_0$  are statistically independent of each other. Furthermore, assume that  $\{w_k\}$  and  $\{v_k\}$  are independent sequences with zero mean

and variance  $I$  and  $\Gamma$ , respectively, and  $x_0$  is a random variable with mean  $\bar{x}_0$  and variance  $\epsilon_0$ . We assume that the following conditions hold for system (Eq. 1):

(1)  $\hat{A}_k = \sum_{i=1}^m \hat{a}_i A_i(k)$ , where  $\hat{\theta}_a = (\hat{a}_1, \dots, \hat{a}_m)^T$  is an unknown  $m \times 1$  constant parameter and  $\{A_i(k), i = 1, 2, \dots, m\}$  are all uniformly bounded matrices.

(2) The state system (Eq. 1) is bibo stable.

(3)  $\lim_{N \rightarrow \infty} \inf \frac{1}{N} \sum_{k=0}^{N-1} F_k > 0$  (positive definite)

where

$$F_k = \begin{bmatrix} \text{tr. } A_1^T(k) A_1(k), & \dots, & \text{tr. } A_1^T(k) A_m(k) \\ \vdots & & \vdots \\ \text{tr. } A_m^T(k) A_1(k), & \dots, & \text{tr. } A_m^T(k) A_m(k) \end{bmatrix}$$

(4)  $E\|w_k\|^8$ ,  $E\|v_k\|^4$ ,  $E\|x_0\|^4$  are all uniformly bounded in  $k$ .

(5)  $\hat{B}_k$  is nonsingular for all  $k$  and is parameterized as  $B_k = \hat{B} E_k$  where  $\hat{B} = (\hat{b}_{ij})$ ,  $i, j = 1, 2, \dots, n$  is an unknown constant matrix and  $E_k$  is a uniformly bounded  $n \times n$  matrix and satisfies

$$D = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} E_k E_k^T$$

exists and is positive definite.

Under the conditions (1) to (5), we first choose a functional  $q_N^{(1)}(\theta_a)$  given by

$$q_N^{(1)}(\theta_a) = \frac{1}{N} \sum_{k=0}^{N-1} \left[ (y_{k+1} - A_k y_k)^T (y_{k+1} - A_k y_k) - \text{tr. } \Gamma - \text{tr. } A_k^T A_k \Gamma \right],$$

then  $q_N^{(1)}(\theta_a)$  satisfies the conditions (1) to (3) given in Theorem 1. Hence we have that the estimate  $\hat{\theta}_a^N$  of  $\hat{\theta}_a$  converges to  $\hat{\theta}_a$  w.p. 1 as  $N \rightarrow \infty$ .

Next we choose a functional  $q_N^{(2)}(\theta_b)$  as

$$q_N^{(2)}(\theta_b) = \left\| \frac{1}{N} \sum_{k=0}^{N-1} \left[ (y_{k+1} y_{k+1}^T - \Gamma - B_k B_k^T) + (\hat{A}_k \Gamma \hat{A}_k^T - \hat{A}_k Y_k Y_k^T \hat{A}_k) \right] \right\|$$

and establish convergence of the estimates  $\hat{\theta}_b^N$  of the magnitudes of the unknown parameters in  $\hat{B}$ .

National Science Foundation  
ATA-7303217

F. Nakijima and F. Kozin

#### REFERENCES

1. P. E. Caines and J. Rissanen, "Maximum Likelihood Estimation for Multivariable Gaussian Stochastic Processes," IEEE, IT-20, No. 1 (1974).
2. T. S. Lee and F. Kozin, "Consistency of Maximum Likelihood Estimate for a Class of Discrete Multivariate Time Varying Models," Proc. 9th Hawaii Intl. Conf. on System Science, Honolulu, Hawaii, (January 1976).
3. L. Ljung, "Consistency of the Least Square Identification Method," IEEE, AC-21, No. 5, (October 1976).
4. L. Ljung, "On Consistency and Identifiability," Proc. Intl. Symp. Stochastic Systems, Lexington, Kentucky (June 1975).
5. L. Ljung, "On the Consistency of Prediction Error Identification Methods," System Identification: Advances and Case Studies. (D. G. Lainiotis and R. K. Mehra, Eds.) Mathematics in Science and Engineering: Vol. 126, Academic Press (1976).
6. F. Nakajima, "Estimation and Modeling for Non-Stationary Time Series," Ph.D. Thesis, Polytech. Inst. of New York (1978).

## A NONLINEAR SERVO FOR LINEAR SYSTEMS

L. Shaw and H. Gambe

This report describes continuing work on the nonlinear control of linear systems<sup>5</sup>. Here, the nonlinear regulator ideas introduced in Ref. 1 are combined with the suggestion in Ref. 2 for the conversion of a regulator controller into a servo controller. The horizon-based servos described here seem to be superior to the nonlinear servos in Ref. 2 with regard to quality of performance (e. g., less overshoot for the same rise-time.)

Although horizon-based controllers require extensive on-line computation, microprocessor advances are markedly reducing that disadvantage.

The use of state observers is also explored, since horizon-based controllers are inherently of state feedback form.

A. Introduction

We will begin with a formulation of the servo configuration in which a nonlinear state regulator can be used to speed the tracking response. Some results about nonlinear horizon-based regulators will then be summarized. Finally, a few simple numerical examples will demonstrate the qualities of these controllers.

Methods are well-known for designing linear feedback servos to control the outputs of linear plants. Here we consider augmentations of such a linear controller by a nonlinear loop which speeds the response when tracking errors are large.

The nonlinear controllers considered here are those previously described for regulation of multivariable linear plants,<sup>1,3</sup> and they can be viewed as providing state-dependent feedback gains.

Since the usual servo is limited to operation on the scalar tracking error (reference minus output), the use of a state-dependent nonlinear controller will generally require use of a state observer to obtain the error-state vector which is to be driven toward zero. Even in the rare cases where all states of the plant are directly observable, direct measurement of the error-state vector will be possible only when tracking step changes in the reference signal. When tracking higher order polynomial reference signals, derivatives of the reference signal (which are not directly measurable) are essential for determination of the error-state vector.

Figure 1 shows the usual linear servo configuration in which  $P(s)$  may include both the given plant and a tandem compensator. If the denominator of  $P(s)$  (of the form  $s^n + a_1 s^{n-1} \dots a_n$ ) has  $a_n = a_{n-1} \dots a_{n-m+1} = 0$ , then  $P$  is a type- $m$  system which

can follow an  $(m-1)$  degree polynomial reference signal  $(r(t) = \sum_{i=1}^m \alpha_i t^{i-1})$  with zero steady-state error ( $\epsilon(t) \rightarrow 0$ ). This kind of single-input, single-output system can be described by the state variable differential equations

$$\dot{\underline{x}} = A \underline{x} + \underline{b} ; \quad y = \underline{c}' \underline{x} \quad (1)$$

where

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & \dots & \dots \\ 0 & 0 & 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & 1 \\ 0 & 0 & -a_{n-m} & \dots & \dots & -a_1 \end{pmatrix} ; \quad \underline{b} = \begin{pmatrix} b_1 \\ \dots \\ \dots \\ \dots \\ b_n \end{pmatrix} ; \quad \underline{c} = \begin{pmatrix} 1 \\ 0 \\ \dots \\ \dots \\ 0 \end{pmatrix} . \quad (2)$$

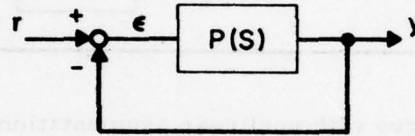


Fig. 1. Feedback servo configuration.

It is convenient to define a reference vector  $\underline{r}(t)$  and an error-state vector  $\underline{e}(t)$  as follows

$$\underline{r} = \begin{pmatrix} r \\ \dot{r} \\ \ddots \\ r^{(n)} \end{pmatrix} ; \quad \underline{e} = \underline{r} - \underline{x} . \quad (3)$$

It follows that the last  $(n-m)$  elements of  $\underline{r}$  are zero and that  $\underline{c}' \underline{e} = \epsilon$ , the scalar tracking error. Moreover, the special canonical form of  $A$  used in Eq. (2) implies

$$A \underline{r} = \dot{\underline{r}} . \quad (4)$$

Substitution of Eqs. (3) and (4) into Eq. (1) yields the differential equations

$$\begin{aligned} \dot{\underline{e}} &= (A - \underline{b} \underline{c}') \underline{e} \\ \underline{e} &= \bar{A} \underline{e} \end{aligned} \quad (5)$$

so that  $\underline{e} \rightarrow 0$  if  $\bar{A}$  is a stable matrix.

Figure 2 shows how the foregoing linear servo can be augmented by an additional nonlinear state regulator in order to effect a faster tracking response. This configuration postulates the use of a dynamic state observer which has knowledge of the first component of  $\underline{e}$  and the nonlinear control input  $u_{NL}(\hat{\underline{e}})$ . The presence of the nonlinear loop changes the error-state vector equations to

$$\dot{\underline{e}} = \bar{A} \underline{e} + \underline{b} u_{NL}(\hat{\underline{e}}). \quad (6)$$

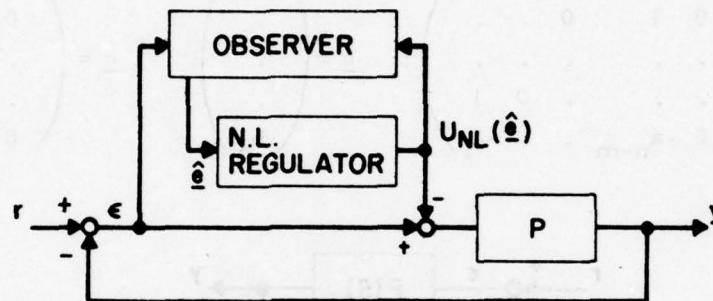


Fig. 2. Servo with nonlinear augmentation loop.

If  $\hat{\underline{e}}$  were to equal  $\underline{e}$  then Eq. (6) would be in the form of a state variable regulator problem for which horizon-based nonlinear controllers have been developed in order to achieve fast reduction of "large" initial  $\underline{e}$ -vectors. If a compatible observer is used,  $\hat{\underline{e}} \rightarrow \underline{e}$ , so that use of Eq. (6) with  $u_{NL}$  designed assuming  $\underline{e} = \hat{\underline{e}}$  will lead to an asymptotically stable servo for which  $\epsilon = e_1 \rightarrow 0$  faster for "large"  $\underline{e}$ . The transient response of this configuration can be studied via simulations.

#### B. Review of Regulator Results

References 1 and 3 describe nonlinear regulators for systems of the form

$$\dot{\underline{e}} = \bar{A} \underline{e} + \underline{B} u(\underline{e}) \quad (7)$$

with  $(\bar{A}, \underline{B})$  controllable. Those regulators are based on linear regulators of the receding horizon or virtual horizon type in which the horizon parameter acts like a scalar measure of the response time. The nonlinear versions make the horizons state-dependent, with "large" error states inducing nearer horizons and correspondingly faster reduction of the error state toward zero.

For example, the linear receding horizon controller takes the form

$$\underline{u}(t) = -R^{-1} \underline{B}' W^{-1}(T) \underline{e}(t), \quad (8)$$

where  $R$  is a positive definite matrix, and  $W(T)$  is the positive definite solution of the linear matrix equation

$$\frac{dW}{dT} = BR^{-1}B' - \bar{A}W - W\bar{A}' ; W(0) = 0. \quad (9)$$

At each instant of time  $t$ , the controller in Eq. (8) acts to drive  $\underline{e}$  to  $\underline{0}$  at a time  $(t+T)$ . Choice of a smaller horizon parameter  $T$  results in a system which regulates faster, while requiring correspondingly larger control energy. (The importance of conserving control energy can be introduced by the choice of  $R$ .)

The nonlinear receding horizon controller makes the horizon distance a function of the state. It is possible to make this  $T(\underline{e})$  a decreasing function of the "size" of  $\underline{e}$  in such a way that the resulting nonlinear system is asymptotically stable, and, as demonstrated by simulations, the system responds faster to larger initial  $\underline{e}(0)$ .

The state-dependent  $T(\underline{e})$  measures the size of  $\underline{e}$  via a quadratic form which is natural for the system, namely

$$V(\underline{e}, T) = \underline{e}'W^{-1}(T)\underline{e} \quad (10)$$

It can be shown that this positive definite function is decreasing in  $T$  (for fixed  $\underline{e}$ ). Thus the described  $T(\underline{e})$  relation is achieved when  $T$  is defined implicitly by the constraint

$$V(\underline{e}, T) = F(T) \quad (11)$$

for a suitable decreasing function  $F(T)$ . (See Figure 3.)

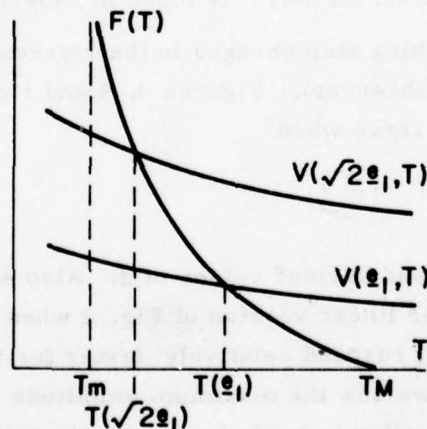


Fig. 3. Implicit definition of  $T(\underline{e})$ .

It can be shown that if

$$\frac{dF}{dT} < \frac{dV}{dT} \quad (12)$$

whenever Eq. (11) is satisfied (i. e., at the intersection points in Fig. 3) then the  $V$  in Eq. (10) is a Lyapunov function, and the closed loop nonlinear system is asymptotically stable. The references discuss methods for finding suitable  $F(T)$  functions which ensure Equation (12). It is clear that if  $F(T)$  decreases monotonically from infinity to zero as  $T$  increases from  $T_m$  to  $T_M$ , then at least one solution ( $T_m < T \leq T_M$ ) to Eq. (11) exists for each  $\underline{e}$ , and the one for which  $T$  is largest will satisfy the stability condition Equation (12).

### C. Examples

Since analytical solution of Eq. (9) is difficult except for system of low dimension, numerical techniques will be most appropriate for finding  $W(T)$  and solving Eq. (11) for  $T(\underline{e})$  at each instant of time. While development of such programs is in progress, some very simple examples have been simulated using analytical solutions of Equation (9).

Referring to Eqs. (2) and (5), a "type 1" plant defined by

$$A = \begin{pmatrix} 0 & 1 \\ 0 & -5 \end{pmatrix} \quad \underline{b} = \begin{pmatrix} 0 \\ 4 \end{pmatrix} \quad \underline{c} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad R = 1$$

results in

$$\bar{A} = \begin{pmatrix} 0 & 1 \\ -4 & -5 \end{pmatrix}$$

for which an explicit expression for  $W(T)$  is found in References 1 and 3.

We first consider tracking step changes in the reference input when the plant states can be measured (no observer). Figures 4, 5 and 6 demonstrate step response properties of this nonlinear servo when

$$F(T) = d^2 \left( \frac{1-T}{T-0.2} \right)^2$$

for various step amplitudes and various values of  $d$ . Also shown are responses of the linear servo of Fig. 1 and the linear version of Fig. 2 when  $T=T_{\max} = 1$ . The ability of the nonlinear controller to respond relatively faster for larger step amplitudes is quite evident. Figure 6 shows how the maximum amplitude of the nonlinear control signal increases as the controller is made "more nonlinear."

Since the tracking of higher order polynomial reference signals will require the use of a state observer, it is of interest to examine this same step response example

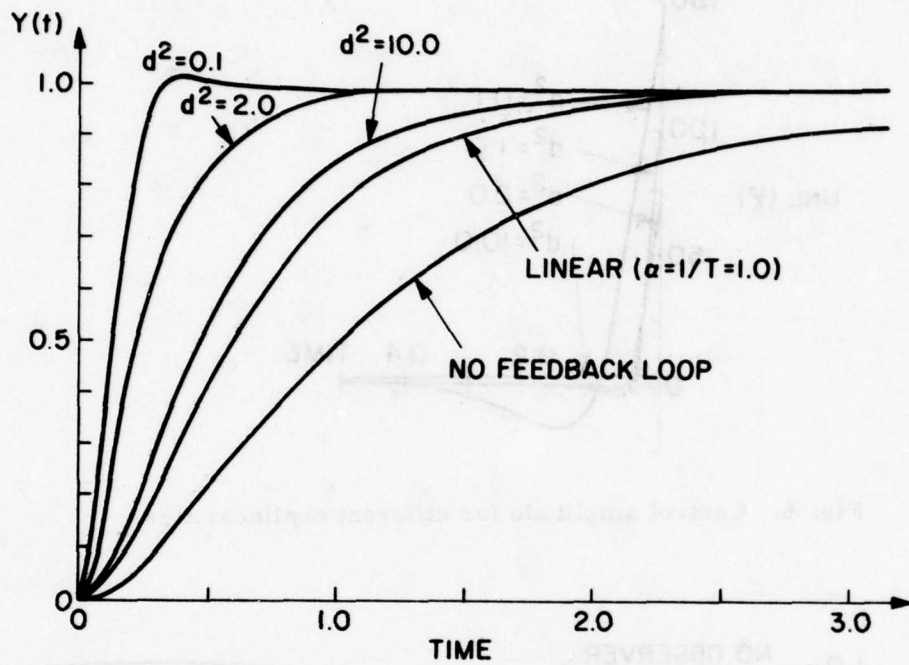


Fig. 4. Different nonlinear step responses.

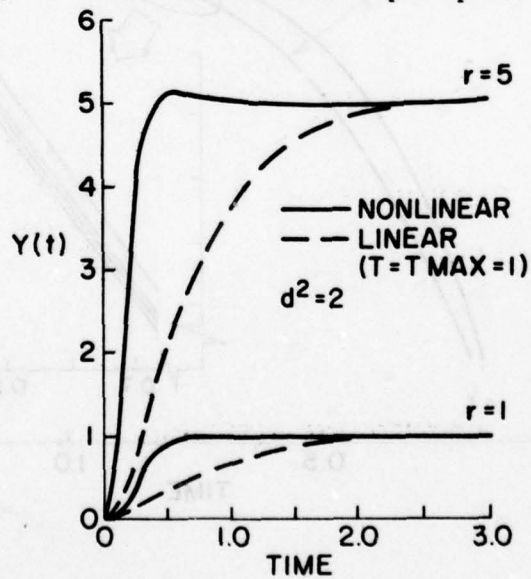


Fig. 5. Nonlinear step response.

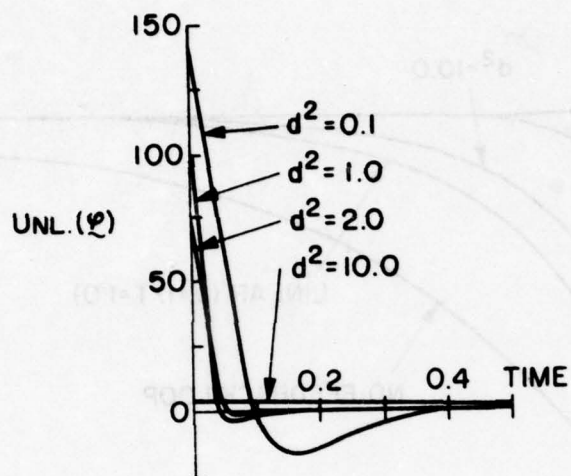


Fig. 6. Control amplitude for different nonlinearities.

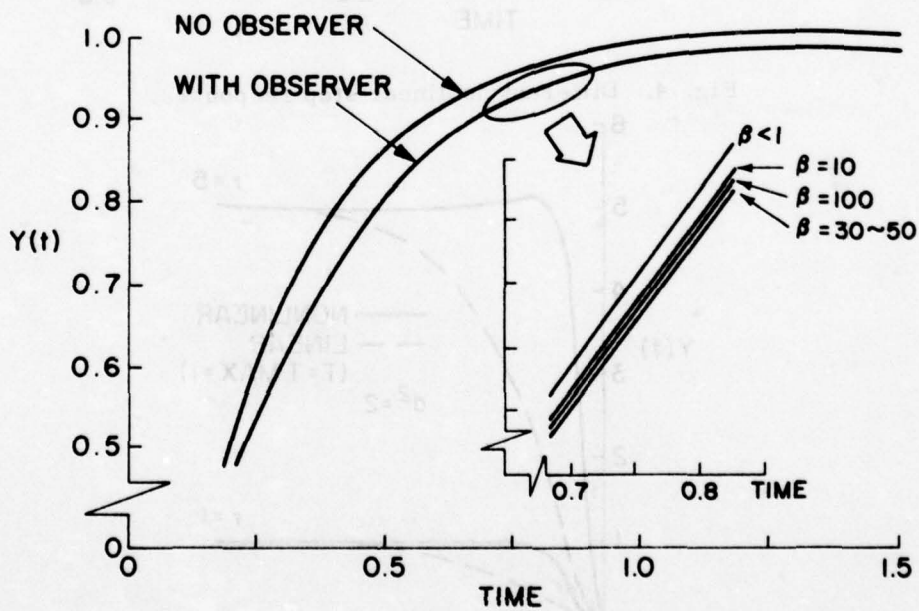


Fig. 7. Nonlinear step response with observer.

when an observer is used. With the observer output  $\hat{\underline{e}}$  defined by

$$\dot{\underline{e}} = F \underline{e} + \underline{h} \epsilon + \underline{b} u$$

we can choose  $F = (A - \underline{h} \underline{c}')$  to have suitably negative eigenvalues to provide fast convergence of the observer error  $\underline{e} = (\underline{e} - \hat{\underline{e}})$ , which satisfies the differential equation:

$$\dot{\underline{e}} = F \underline{e}.$$

(The full-dimension observer has been used because it has been noted that in some control problems this causes less performance degradation than the reduced order observer.<sup>4</sup>)

The effects of observer convergence rate on the behavior of the nonlinear controller were investigated by choosing

$$|sI - F| = (s + \beta)^2$$

for several values of  $\beta$ . Figure 7 shows the unusual results. The step response is more sluggish here, as compared to the perfect state observation case, for all  $\beta$ . As  $\beta$  increases, from 1 to 30 to 50, the response becomes slower and slower; but changing from  $\beta = 50$  to  $\beta = 150$  improves the speed of response!

#### D. Conclusion

It has been shown that a nonlinear, horizon-based regulation can be used to augment the tracking performance of a linear servo; more work is needed to examine the interrelationship of the observer convergence to a precise estimate of the error-state vector, and the servo convergence of the error-state vector to zero.

Joint Services Electronics Program  
F44620-74-C-0056

L. Shaw, H. Gambe

#### REFERENCES

1. L. Shaw, D. Scarlat, and Y. Thomas, "Synthesis of Nonlinear Controllers," IFAC, 7th World Congress, Helsinki (June 1978)
2. J. Sandor and D. Williamson, "Nonlinear Feedback to Improve the Transient Response of a Linear Servo" IEEE Trans. on Automatic Control, Vol. 22, pp. 863-864 (October 1977.).
3. L. Shaw, "Nonlinear Control of Linear Multivariable Systems via State Dependent Feedback Gains," 1978 Conference on Decision and Control, San Diego (January 1979).
4. J. J. Bongiorno, Jr., and D. C. Youla, "On Observers in Multi-Variable Control Systems," Int. J. Control Vol. 8, No. 3, pp. 221-243.
5. L. Shaw, "Nonlinear Control of Linear Multivariable Systems," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452. 42-77 (1977).

## EFFICIENT DISCRETE FOURIER TRANSFORMATION OF REAL VECTORS

T. W. Parsons

Using results obtained from the study of fast convolution algorithms, S. Winograd of IBM has found a discrete Fourier transform (DFT) algorithm<sup>1</sup> which is more efficient than the well-known Cooley-Tukey ("FFT") algorithm. Efficiency in the DFT is customarily measured by the number of multiplications required, on the assumption that multiplications are the most time-consuming operations and that therefore they account for most of the computation time required. Where the FFT requires  $O(N \log_2 N)$  real multiplications to transform a complex  $N$ -vector, Winograd's algorithm requires  $O(2N)$  real multiplications.

In many applications of the DFT, the vector to be transformed is real; in that case, half of the resulting transform is redundant, according to the well-known property of such transforms,

$$Y(N-n) = \overline{Y(n)} \quad , \quad (1)$$

and therefore a certain amount of unnecessary computation is done. In spite of its efficiency, Winograd's algorithm, while its structure lays bare the symmetries inherent in the transforms of real vectors, nevertheless ends up carrying the usual overhead of redundant values. We present here an adaptation of Winograd's algorithm which requires no complex arithmetic, except at the point where the transform vector is generated or (in the inverse transform) accepted, and which does not compute any redundant values.

We write the DFT of an  $N$ -vector as follows:

$$Y = W y \quad , \quad (2)$$

where  $W$  is an  $N \times N$  matrix whose  $(n, k)$  element is  $w_N^{nk}$  and  $w_N = \exp(2\pi j/N)$ . Indexing is zero-origin throughout and the exponent is computed modulo  $N$ .

For  $N$  a prime or a power of a prime, Winograd's algorithm can be represented as a factorization of the  $W$  matrix,

$$W = S C T \quad . \quad (3)$$

In this factorization, the multiplications are confined to  $C$ , a diagonal matrix of approximately  $N$  multipliers; the  $S$  and  $T$  matrices consist only of 0's and  $\pm 1$ 's and thus merely implement output and input additions (and subtractions) of the elements. Such factorizations are known, and set forth in Ref. 1, for  $N \in \{2, 3, 4, 5, 7, 8, 9, 16\}$ . Because of their limited size, these factorizations are known as "small- $N$ " transforms. The structures of all of these factorizations are similar; for this summary we will take

the case,  $N=5$ , as representative. For  $N=5$ ,

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & -1 & 0 \\ 1 & 1 & -1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & 0 & -1 \\ 1 & 1 & 1 & -1 & 1 & 0 \end{bmatrix} \quad (4)$$

$$C = \text{diag}(1, -1.25, .559, j.951, j1.54, -j.363) \quad (5)$$

$$T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & -1 & 1 \\ 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix} \quad (6)$$

Note that the elements of  $C$  are always either pure real or pure imaginary.

If it is known that  $y$  is real, then the elements of  $Y$  satisfy Equation (1). It is not difficult to rewrite  $S$ ,  $C$  and  $T$  so as to separate the real and imaginary parts of  $Y$  and place the imaginary parts in the locations otherwise used for the redundant elements. Formally, we may define a real vector  $H$  such that

$$H(n) = \begin{cases} \text{Re}[Y(n)] & n \leq N/2 \\ \text{Im}[Y(N-n)] & \text{otherwise} \end{cases} \quad (7)$$

The elements of  $Y$  can be easily recovered from  $H$  as follows:

$$Y(n) = \begin{cases} H(n) & N = 0, N/2 \\ H(n) + jH(N-n) & 0 < N < N/2 \\ H(N-n) - jH(n) & N/2 < N < N \end{cases} \quad (8a)$$

or

$$Y = BH \quad (8b)$$

where the elements of  $B$  are readily determined from inspection of Equation (8a). Then we may write,

$$H = S' C' T y \quad (9)$$

where  $S'$  and  $C'$  are modifications of the  $S$  and  $C$  matrices of Winograd, and

$$Y = BS' C' T y \quad (10)$$

The modified matrices are found as follows. For  $S'$ , it is a matter of simple inspection to identify those portions of  $S$  which contribute, on the one hand, to the real and imaginary parts of  $Y$  and, on the other, to the  $n=0$ ,  $0 < n < N/2$ ,  $n=N/2$ , and  $N/2 < n < N$  elements of  $Y$ . Doing so leads to a partitioning of the  $S$  matrix, and to form  $S'$  from  $S$ , we need only drop the redundant partitions. For example, for  $N=5$ ,

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & -1 & 0 \\ 1 & 1 & -1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & 0 & -1 \\ 1 & 1 & 1 & -1 & 1 & 0 \end{bmatrix} \quad (11)$$

$$S' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} \quad (12)$$

Finally,  $C'$  is formed from  $C$  by multiplying its imaginary elements by  $-j$ .

In computing the DFT of a real vector using the factorization Eq. (10), all arithmetic is real until the final multiplication by  $B$ . As can be seen from Eq. (8), however, multiplication by  $B$  consists merely of assembling the elements of  $Y$  from those of  $H$ ; hence we have for computational purposes eliminated all complex arithmetic.

For  $N$  large and composite, a more elaborate process is required, both for Winograd's algorithm and for our modification. It has been shown by Thomas<sup>3</sup> that if  $N$  can be factored into  $\ell$  mutually prime factors  $N_1, N_2, \dots, N_\ell$ , then the 1-dimensional  $N$ -point DFT can be replaced by an equivalent  $\ell$ -dimensional DFT each of whose dimensions corresponds to one of the factors of  $N$ . In the Winograd algorithm, each dimension is transformed using a small- $N$  transform factored as in Eq. (2); because of linearity, the various steps for the various dimensions can be handled in any convenient order, and in Winograd's algorithm all the  $T$ -multiplications are done first, then all the  $C$ -multiplications, etc.

When this procedure is applied to real vectors using Eq. (1), we see that again all arithmetic required is real until the final multiplication of each dimension by its proper  $B$  matrix. To find a way to implement the  $B$ -multiplications without requiring extensive storage of complex intermediate values takes some ingenuity. Let  $\underline{Y}$  be the  $\ell$ -dimensional working array just before the final multiplications by the  $B$ -matrices, and let  $Z$  be the  $\ell$ -dimensional output transform. For example, for  $\ell = 2$ ,

$$Z = B_1 \underline{Y} B_2^t \quad (13)$$

To combine these multiplications into a single step, we introduce the stacking operator  $v(\cdot)$ , defined as follows: for any array  $A$ ,  $v(A)$  is the column vector formed by stacking the columns of  $A$ . It can easily be shown<sup>4</sup> that using the stacking operator, Eq. (13) can be rewritten

$$v(Z) = (B_1 \times B_2) v(\underline{Y}) \quad (14)$$

where  $B_1 \times B_2$  is the Kronecker product of  $B_1$  and  $B_2$ . (Although Eq. (14) applies to

two dimensions only, it can be generalized to  $l$  dimensions by induction on the number of dimensions.)

Using the form Eq. (14), we can combine all of the final B-matrix multiplications into a single step with a real input vector  $v(\underline{X})$ , a complex output vector  $v(Z)$ , and no complex storage of intermediate results other than a simple accumulator.

Space does not permit describing the implementation of Eq. (14) in the computer, but we may observe, first, that the Kronecker product  $(B_1 \times B_2 \times \cdots \times B_l)$  does not have to be stored in memory; its elements can be computed as required from the elements of the B-matrices, which in turn can be found as needed from the definition Equation (8). Second, note that multidimensional arrays are customarily stored in computer memory in stacked form and that the indices required to single out an element of are also those needed to address the corresponding elements of the constituent B matrices. This leads naturally to a nested-loop program structure. Finally, by using nested loops and bypassing all inner levels of nesting whenever a B-matrix element is found to be zero, we can reduce the  $N^2$  multiplications implicit in the definition of  $(B_1 \times B_2 \times \cdots \times B_l)$  to somewhat fewer than  $2^l N$  additions. A more detailed description will be found in Reference 5.

The inverse transform is found by taking the conjugate transpose of  $W$ ; this reverses the order of the factorization in Eq. (10):

$$W^{-1} = \frac{1}{N_i} W^* = \frac{1}{N_i} T^t C^t S^t B^* \quad (15)$$

In the multidimensional case, the  $B^*$  multiplications now come first and can be implemented by means of the Kronecker-product form as in the direct transform.

EE Departmental Research

T. Parsons

#### REFERENCES

1. S. Winograd, "On Computing the Discrete Fourier Transform," IBM Research Report RC-6291 (November 1976).
2. J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Computation of Complex Fourier Series," Math. of Comp., Vol. 19, No. 90, pp. 297-301 (April 1965).
3. L. H. Thomas, "Using a Computer to Solve Problems in Physics," in Applications of Digital Computers, Boston: Ginn (1963).
4. D. H. Nissen, "A Note on the Variance of a Matrix," Econometrica, Vol. 36, No. 3-4, pp. 603-604 (July-October 1968).
5. T. Parsons, "A Winograd Fourier Transform Algorithm for Real-Valued Data," (submitted to IEEE Trans. on Acous., Speech and Sig. Processing).

## SOME COMMENTS ON IMPROVING THE EFFICIENCY OF LEAST SQUARES ESTIMATION

I. Kadar and L. Kurz

It is well known that least-squares estimators are inefficient when pdf of the noise is heavy tailed. The efficiency can be improved, however, by a batch-nonlinear-integer rank transformation which guarantees asymptotic normality after  $m$ -batch pre-processing steps ( $m \approx 10$ ) and robustizes the least-squares estimators.

Consider the relative performance of the least-squares estimator,  $b$ , and the best linear unbiased estimator,  $\hat{\beta}$  of  $\beta$  in the linear model

$$y = X\beta + V$$

where  $y$ , which is  $n \times 1$  and  $X$  which is  $n \times r$  are observed and the  $n \times 1$  noise (error) vector has  $E(V) = 0$ ,  $E[VV^T] = \text{Var}V = \Gamma$ . Assume that the rank  $(X) = r$ , rank  $(\Gamma) = n$  and  $n \geq r$ . Then it is well-known<sup>1</sup> that

$$b = (X^T X)^{-1} X^T y, \quad \hat{\beta} = (X^T \Gamma^{-1} X)^{-1} X^T \Gamma^{-1} y$$

$$\text{Var}(b) = (X^T X)^{-1} X^T \Gamma X (X^T X)^{-1}, \quad \text{Var}(\hat{\beta}) = (X^T \Gamma^{-1} X)^{-1}$$

and  $\text{Var}(b) - \text{Var}(\hat{\beta})$  is non-negative definite. There is no loss in generality in supposing that

$$X^T X = [I]_{r \times r}$$

since  $X$  is given.

The question arises as to how good or how bad is  $b$  relative to  $\hat{\beta}$ ? One needs a relative measure of performance. Since  $X$  is assumed known, but  $\Gamma$  is known rarely, it is of interest to compute a lower bound to the efficiency measure for  $X$  fixed and all  $\Gamma$  in some class. One possible measure suggested by Bloomfield and Watson<sup>2</sup> is the ratio of generalized variances. The efficiency of  $b$  relative to  $\hat{\beta}$  is then

$$\text{Eff} = \frac{|\text{Var}(\hat{\beta})|}{|\text{Var}(b)|} = \frac{\{ |X^T \Gamma X| |X^T \Gamma^{-1}| \}}{\{ |X^T \Gamma X| |X^T \Gamma^{-1}| \}} \leq 1$$

where,  $|\cdot|$  is the determinant of the matrix.

Conditions for which  $\text{Eff}$  attains unity are well-known. If  $\Gamma$  is  $\sigma^2[I]$ , (identity matrix), derived from an i.i.d population, then  $\text{Eff} = 1$ . This suggests, directly, the use of a recursive formulation of the above problem via robustized stochastic approximation minimum variance least squares (SAMVLS)<sup>3</sup> assuming, of course, that multiple

observations are available where we asymptotically attain  $\text{Eff} = 1$ . This means that having formed the optimum gain matrix or an appropriate estimator thereof, asymptotically we obtain a robust estimator for each component  $\beta_{lk} \rightarrow \hat{\beta}_l$  and then

$$V_{SK_{ll}} \geq \text{Var}(V_{ii})/n \alpha_{ii}^2, \quad i = 1, 2, \dots, r$$

where  $V_{SA_{ll}}$  are the diagonal elements of  $\Gamma$  and  $\hat{\beta}_l$  is the best linear unbiased estimator and  $V_{SA_{ll}}$  is independent of the underlying CDF.

Once robustness is guaranteed (i.e.  $b = \hat{\beta}$ ) we are now at liberty to choose a better design matrix,  $X$ , for a particular experiment. This is the subject of optimal experiments,<sup>4</sup> where it is assumed a priori that  $b = \hat{\beta}$ . Therefore, we have extended the theory of least squares by the introduction of robustness, as applied to optimal designs.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

I. Kadar and L. Kurz

Grumman Aerospace Corporation

#### REFERENCES

1. A. P. Sage and J. L. Melsa, "Estimation Theory with Applications to Communications and Control," McGraw-Hill, 1971.
2. P. Bloomfield and G. S. Watson, "The Inefficiency of Least Squares," *Biometrika*, Vol. 62, No. 1, April 1975.
3. I. Kadar and L. Kurz, "Robustized Scalar Form of Gladyshev's Theorem with Applications to Nonlinear Systems," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77, pp. 488-499 (1977).
4. V. V. Fedorov, "Theory of Optimal Experiments," Academic Press, 1972.

## ADAPTIVE FREQUENCY DOMAIN ESTIMATORS

A. Papoulis

A new method of estimation is presented combining frequency domain techniques with the advantages of adaptive and recursive filtering. The method is based on the representation of an arbitrary signal in terms of the samples  $F(t, m\omega_0)$  of its "running Fourier transform"  $F(t, \omega)$  and it assumes no prior knowledge of any statistics.

The algorithm for determining  $F(t, m\omega_0)$  is a first order recursion. The filter weights are determined adaptively and the adaptation algorithm involves either a running estimate of the required statistics or an instantaneous version of a gradient-seeking LMS search leading to a generalization of the Widrow filter.

The study includes the development of various properties of running transforms and their use in spectral estimation.

A. Adaptive and Recursive MS Estimators

Adaptive filters are useful because they require no knowledge of prior statistics and they can be used to process stationary and non-stationary signals.<sup>1,2</sup> Recursion is introduced to simplify the required arithmetic operations. In this report, we develop a new filter that combines these advantages with the advantages of frequency domain processing techniques. To place our development into familiar perspective, we discuss first, with some originality, two known methods of adaptive filtering. We shall use as illustration the estimation of a discrete process  $y[n]$  in terms of the output

$$\hat{y}[n] = \sum_{k=0}^{N-1} a_k[n] x[n-k] = X^T[n] A[n] \quad (1)$$

of a non-recursive, time-varying system with input the  $N$  samples  $x[n-k]$  of the data vector  $X[n]$ . Our objective is to determine the  $N$  components  $a_k[n]$  of the transfer vector  $A[n]$  so as to minimize in some sense the estimation error

$$e[n] = y[n] - \hat{y}[n] = y[n] - X^T[n] A[n] \quad (2)$$

1. The Adaptive Wiener Filter

If the processes  $x[n]$  and  $y[n]$  are jointly stationary and the optimality criterion is the minimization of the MS error  $E\{e^2[n]\}$ , then the transfer vector  $A[n]$  is independent of  $n$  and is found by solving the system

$$R A = \Gamma \quad (3)$$

where

$$R = E\{X[n] X^T[n]\} \quad (4)$$

is the correlation matrix of the data vector and

$$\Gamma = E\{X[n] y[n]\} \quad (5)$$

is the cross-correlation vector between the data and the signal to be estimated.<sup>3</sup> The resulting solution

$$A = R^{-1} \Gamma \quad (6)$$

is, of course, the transfer vector of the Wiener filter. This filter is optimum but it requires prior knowledge of the matrix  $R$  and the vector  $\Gamma$ .

In the absence of this knowledge, we can use for  $A[n]$  the solution of the system

$$\hat{R}[n] A[n] = \hat{\Gamma}[n] \quad (7)$$

where

$$\hat{R}[n] = (1 - \alpha) \sum_{k=0}^n \alpha^k X[n-k] X^T[n-k] \quad (8)$$

$$\hat{\Gamma}[n] = (1 - \alpha) \sum_{k=0}^n \alpha^k X[n-k] y[n-k] \quad (9)$$

and

$$0 < \alpha < 1 \quad (10)$$

It can be shown that the quantities  $\hat{R}[n]$  and  $\hat{\Gamma}[n]$  tend to  $R$  and  $\Gamma$ , respectively, as  $n \rightarrow \infty$  and  $\alpha \rightarrow 1$ . Hence, for sufficiently large  $n$  and for  $\alpha$  close to one (see the solution

$$A[n] = \hat{R}^{-1}[n] \hat{\Gamma}[n] \quad (11)$$

of Eq. (7) is nearly optimum as in Equation (6).

## 2. The Widrow Filter

In the Widrow filter,<sup>4,5</sup> the adaptation algorithm is a first order recursion:

$$A[n+1] = A[n] + \mu e[n] X[n] \quad (12)$$

where  $\mu$  is some constant that determines the time of adaptation. This algorithm is justified as an instantaneous version of the gradient-seeking method for minimizing

the MS estimation error.

The Widrow filter is simple; it requires no prior knowledge of any statistics; it can be applied to stationary and non-stationary processes. It has, however, a number of disadvantages.

The solution  $A[n]$  of Eq. (12) does not approach a constant as  $n \rightarrow \infty$ . The vector  $A[n]$  remains random for any  $n$ . Various attempts have been made to analyze Eq. (12), however, the known results are based on assumptions that are not generally valid.<sup>5,6</sup> The difficulty is that Eq. (12) is a time-varying equation with random coefficients. Indeed, inserting Eq. (2) into Eq. (12), we obtain

$$A[n+1] - (1 - \mu X^T[n] X[n]) A[n] = \mu X[n] y[n] \quad (13)$$

Several investigations are based on the assumptions that

$$E\{X^T[n] X[n] A[n]\} = E\{X^T[n] X[n]\} E\{A[n]\} \quad (14)$$

If this holds, then the mean of  $A[n]$  tends to the Wiener optimum Eq. (6) as  $n \rightarrow \infty$ . This is the usual justification of using Eq. (12) as adaptation algorithm. One of the justifications of Eq. (14) that appears on the surface to be not too restrictive, is the assumption that all processes are normal with zero mean.<sup>6</sup> Whereas the normality assumption for the data is not too restrictive, it does not hold for the process  $A[n]$  because its dependence on  $x[n]$  and  $y[n]$  is strongly nonlinear as we see from Equation (17). The problem of determining the statistical properties of the solution of Eq. (13) has been investigated for various special cases,<sup>7,8</sup> but the results are by no means conclusive. Numerical studies, however, have shown that, in a number of applications, the filter performs well.

### 3. The Adaptive FFT Filter

In the preceding filters, the number of weights  $a_k[n]$  that are adaptively controlled equals the length  $N$  of the filter (see Equation (1)). This often introduces an unnecessary complexity in the design if the required value of  $N$  is large. We mention, as a simple illustration, the adaptive realization of a one-pole system with the filter in Equation (1). In this case, the coefficients  $a_k[n]$  must approach a geometric progression  $z_0^k$  involving a single parameter  $z_0$ . For a satisfactory non-recursive approximation, however, the length  $N$  of the filter must be large if  $z_0$  is close to one.

To overcome this problem we impose the restriction that the  $N$  weights  $a_k[n]$  are linearly dependent on  $M_0$  parameters  $b_m[n]$

$$A[n] = WB[n] \quad (15)$$

In the above,  $B[n]$  is an  $M_0$ -vector with elements  $b_m[n]$  and  $W$  is an  $N$  by  $M_0$  matrix to be determined. Inserting Eq. (1), we obtain

$$\hat{y}[n] = X^T[n]WB[n] = Z^T[n]B[n] \quad (16)$$

where

$$Z[n] = W^T X[n] \quad (17)$$

Equations (16) and (17) specify the system of Fig. 1 consisting of the matrix  $W$  transforming the  $N$ -vector  $X[n]$  into the  $M_0$ -vector  $Z[n]$ , followed by the weights  $B[n]$  and an adder. The  $M_0$ -vector  $B[n]$  is adaptively controlled.

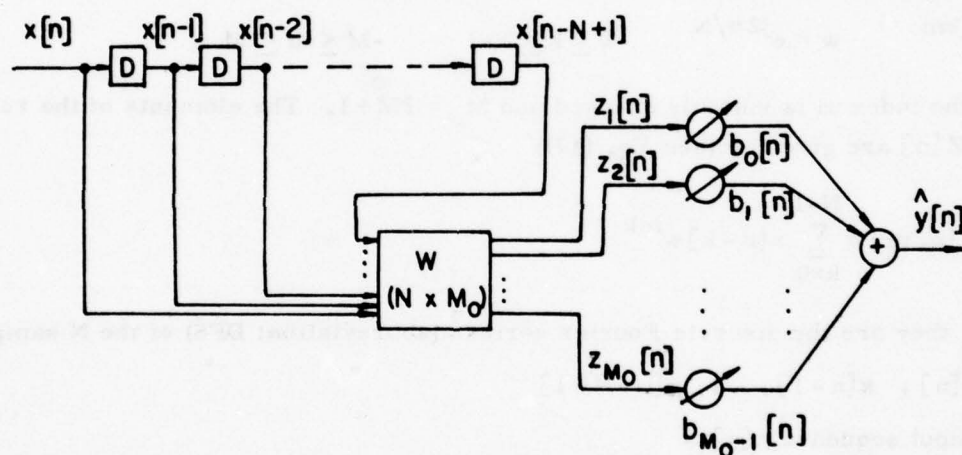


Fig. 1.

The selection of the matrix  $W$  depends on the properties of the processes  $x[n]$  and  $y[n]$ . Suppose, for example, that

$$x[n] = y[n] + v[n] \quad (17)$$

where  $v[n]$  is noise, the vectors  $X[n]$  and  $Y[n]$  are elements of an  $N$ -dimensional space. If the signal vector  $Y[n]$  belongs to a sub-space  $S_y$  of dimensionality  $M_0$ , then we choose for  $W$  the projection operator into  $S_y$ .

Another consideration in the selection of  $W$  is the nature of the resulting vector  $Z[n]$ . If the autocorrelation matrix of  $Z[n]$  is diagonal or nearly so, then the evaluation of the inverse of its sample matrix  $\hat{R}^{-1}[n]$  used in the adaptive Wiener filter (see Eq. (11)) is considerably simplified.

These two considerations, small  $M_0$ ; data vector  $Z[n]$  with uncorrelated components, are satisfied if we use for  $W$  the matrix resulting from the Karhunen-Loeve expansion<sup>3</sup> of  $X[n]$ . The coefficients of this expansion depend, however, on the statistics of  $X[n]$  which we assumed unknown.

In our search for a suitable transformation, we are guided also by another factor: the spectral properties of the various signals. In a number of cases, filtering is best performed in the frequency domain - elimination of high frequency noise, or low frequency clutter, for example. In the design of the filter, it is desirable to select as adaptively controlled variables, parameters that are directly related to the frequency domain characteristics of the various signals.

To meet these requirements, we select as  $W$  a matrix whose elements are the roots of unity:

$$w^{km} \quad w = e^{j2\pi/N} \quad 0 \leq k \leq N-1 \quad -M \leq m \leq M \quad (19)$$

where the index  $m$  is suitably changed and  $M_0 = 2M+1$ . The elements of the resulting vector  $Z[n]$  are given by (see Eq. (17))

$$F[n, m] = \sum_{k=0}^{N-1} x[n-k] w^{mk} \quad (20)$$

that is, they are the discrete Fourier series<sup>9</sup> (abbreviation: DFS) of the  $N$  samples

$$x[n], x[n-1], \dots, x[n-N+1]$$

of the input sequence  $x[n]$ .

With this choice of  $W$ , the output of the filter of Fig. 1 is given by (see Eq. (16))

$$\hat{y}[n] = \sum_{m=-M}^M b_m[n] F[n, m] \quad |m| \leq M \quad (21)$$

and the coefficients  $b_m[n]$  are adaptively controlled.

It appears from Eq. (20) that, for the determination of the  $M_0$  components  $F[n, m]$  of the vector  $Z[n]$ , the required number of multiplications equal  $M_0$  times  $N$ . However, as we show in Section C, each  $F[n, m]$  can be computed recursively with one multiplication only.

In the following, we study various properties and extensions of the Fourier filter of Figure 1. The results are based on the properties of running transforms to be developed next.

### B. Running Fourier Transforms

The running transform  $F(t, \omega)$  of a signal  $f(t)$  is the integral<sup>9</sup>

$$F(t, \omega) = \int_{-c}^c f(t + \tau) e^{-j\omega\tau} d\tau \quad (22)$$

where  $c$  is a given constant. For a fixed  $t$ ,  $F(t, \omega)$  is the Fourier transform in the variable  $\tau$  of the segment  $f(t + \tau)$  of  $f(t)$  shown in Figure 2.

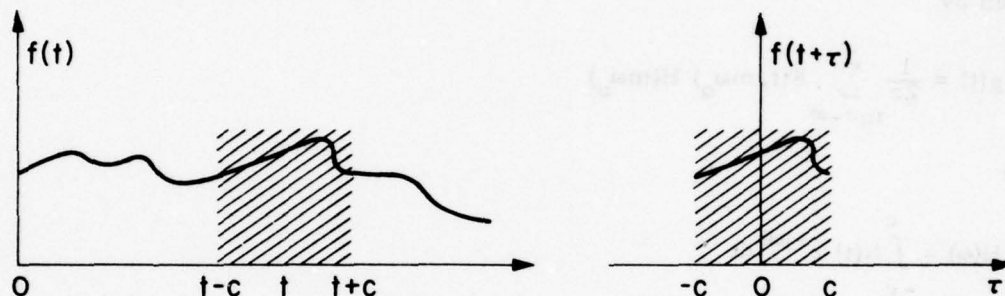


Fig. 2.

We shall presently show that the response  $g(t)$  of any time-limited system can be expressed in terms of the samples  $F(t, m\omega_0)$  of  $F(t, \omega)$ . We start with the following version of the sampling theorem:<sup>9,10</sup>

1. Inversion Formula: For any  $t$ ,

$$f(t) = \frac{1}{2c} \sum_{m=-\infty}^{\infty} F(t, m\omega_0) \quad (23)$$

Proof. We expand the function  $f(t + \tau)$  into a Fourier series in the variable  $\tau$ . The coefficients of the expansion in the interval  $(-c, c)$  are given by (see Eq. (4))

$$\frac{1}{2c} \int_{-c}^c f(t + \tau) e^{-jm\omega_0\tau} d\tau = \frac{1}{2c} F(t, m\omega_0) \quad (24)$$

Hence, for  $|\tau| < c$

$$f(t + \tau) = \frac{1}{2c} \sum_{m=-\infty}^{\infty} F(t, m\omega_0) e^{jm\omega_0\tau} \quad (25)$$

and Eq. (23) results with  $\tau = 0$ .

2. Filtering: We next show that if the impulse response  $h(t)$  of a linear system is time-limited:

$$h(t) = 0 \quad \text{for} \quad |t| > c \quad (26)$$

then its response

$$g(t) = \int_{-c}^c f(t - \alpha) h(\alpha) d\alpha \quad (27)$$

is given by

$$g(t) = \frac{1}{2c} \sum_{m=-\infty}^{\infty} F(t, m\omega_0) H(m\omega_0) \quad (28)$$

where

$$H(\omega) = \int_{-c}^c h(t) e^{-j\omega t} dt \quad (29)$$

Proof. Inserting Eq. (25) into Eq. (27) and integrating term-wise, we obtain Equation (28).

Varying the number of terms in Eq. (28) or the values of the coefficients  $H(m\omega_0)$ , we can design filters whose frequency response is adaptively controlled. We shall elaborate in the development of the discrete version of Equation (28). We note next an extension of the familiar smoothing method used to estimate a signal  $y(t)$  in terms of the data

$$f(t) = y(t) + v(t) \quad (30)$$

containing the noise process  $v(t)$ .

3. Running Low-Pass Filters: Suppose that

$$H(m\omega_0) = \begin{cases} 1 & \text{for } |m| \leq M \\ 0 & \text{for } |m| > M \end{cases} \quad (31)$$

In this case, Eq. (28) yields

$$g(t) = f_M(t) = \frac{1}{2c} \sum_{m=-M}^M F(t, m\omega_0) = \frac{1}{2c} F(t, 0) + \frac{1}{c} \sum_{m=1}^M \operatorname{Re} F(t, m\omega_0) \quad (32)$$

The first term above is the average of  $f(t)$  in the interval  $(-c, c)$ :

$$f_0(t) = \frac{1}{2c} F(t, 0) = \frac{1}{2c} \int_{-c}^c f(t + \tau) d\tau \quad (33)$$

If  $f_0(t)$  is used as the estimate of  $y(t)$ , then the variance of the estimate decreases with increasing  $c$  - it equals  $1/2c$  if the autocorrelation of the noise equals  $\delta(\tau)$ . However, its bias, due to the smoothing of  $y(t)$ , increases. For an optimum estimation for each  $t$ , the length  $2c$  of the smoothing interval  $2c$  must be adaptively varied.<sup>11</sup> As we show next, the effective length of the smoothing interval can be controlled by the number  $M$  in Equation (32).

From Eqs. (22) and (32) it follows that

$$g(t) = \frac{1}{2c} \sum_{m=-M}^M \int_{-c}^c f(t-\alpha) e^{jm\omega_0 \alpha} d\alpha = \frac{1}{2c} \int_{-c}^c f(t-\alpha) \sum_{m=-M}^M e^{jm\omega_0 \alpha} d\alpha$$

This shows that the impulse response  $h(t)$  of the running low-pass filter is given by

$$h(t) = \frac{p_c(t)}{2c} \sum_{m=-M}^M e^{jm\omega_0 t} = p_c(t) \frac{\sin(M + \frac{1}{2})\omega_0 t}{2c \sin \frac{\omega_0 t}{2}} \quad (34)$$

where

$$p_c(t) = \begin{cases} 1 & |t| \leq c \\ 0 & |t| > c \end{cases}$$

The transform of  $h(t)$  yields the corresponding frequency response:

$$H(\omega) = \sum_{m=-M}^M \frac{\sin c(\omega - m\omega_0)}{c(\omega - m\omega_0)} \quad (35)$$

In Fig. 3, we plot  $h(t)$  and  $H(\omega)$  for fixed  $c$  and for  $M$  from zero to five. The effective duration of  $h(t)$  (main lobe) equals

$$t_c = \frac{2\pi}{(2M+1)\omega_0} = \frac{2c}{2M+1} \quad (36)$$

and it decreases with increasing  $M$ . As we see from the figure,  $H(\omega)$  is essentially a low-pass filter even for moderate values of  $M$  and its cut-off frequency equals

$$\omega_c = (M + \frac{1}{2})\omega_0 = \frac{\pi}{t_c} \quad (37)$$

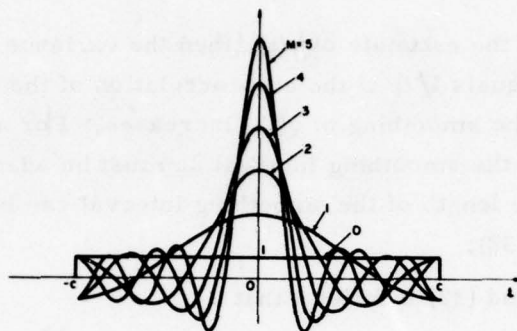


Fig. 3(a).

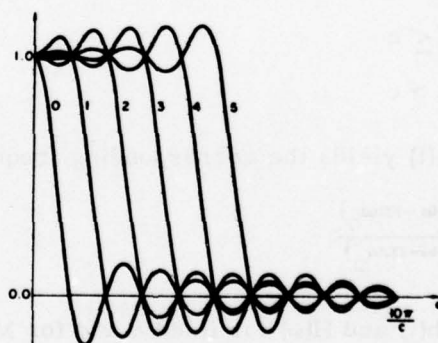


Fig. 3(b).

In Fig. 4, we plot  $h(t)$  and  $H(\omega)$  for a fixed  $t_c$  and for  $M$  from zero to five. We have thus a method for controlling the bandwidth of the filter or the duration of its impulse response by varying  $c$  or  $M$ .

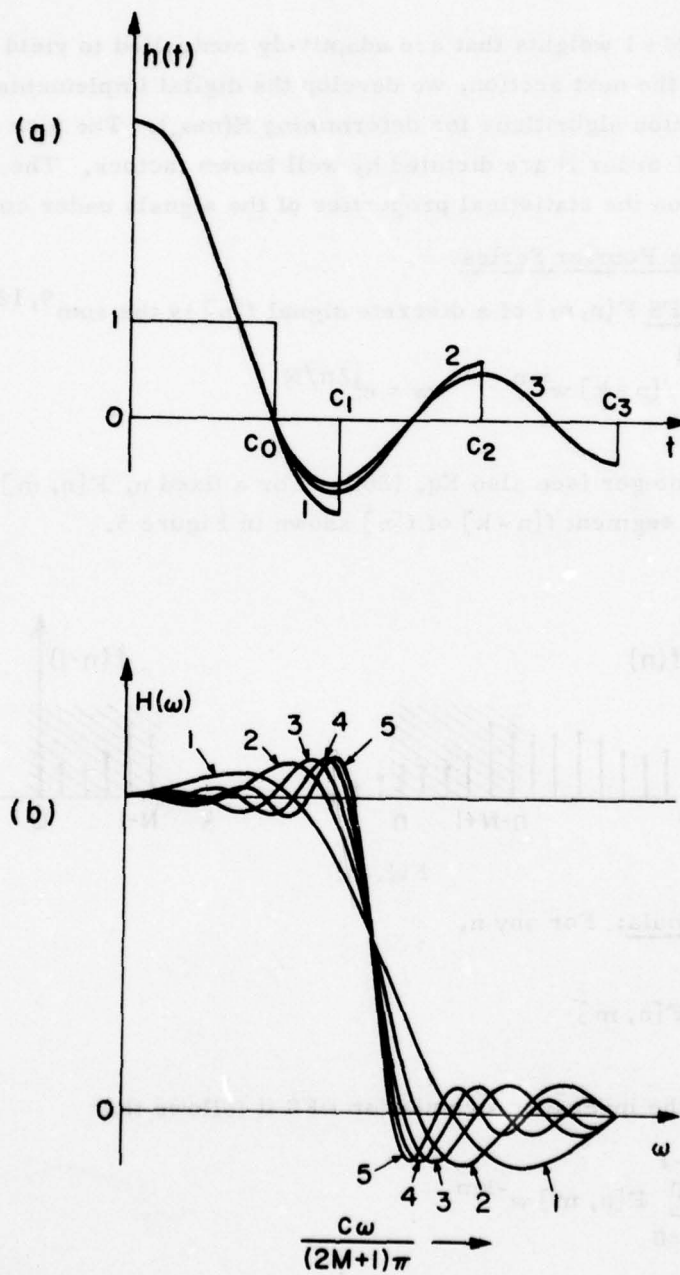


Fig. 4.

The running Fourier filter is an extension of Eq. (32) and it is obtained if the sum in Eq. (28) is suitably truncated. Its response equals the sum

$$\frac{1}{Z_c} \sum_{m=-M}^M F(t, m\omega_0) H(m\omega_0) \quad (38)$$

where  $H(m\omega_0)$  are  $2M+1$  weights that are adaptively controlled to yield estimates of various signals. In the next section, we develop the digital implementation of the filter and the adaptation algorithms for determining  $H(m\omega_0)$ . The size of the sampling interval and the FFT order  $N$  are dictated by well known factors. The value of the constant  $M$  depends on the statistical properties of the signals under consideration.

### C. Running Discrete Fourier Series

The running DFS  $F[n, m]$  of a discrete signal  $f[n]$  is the sum<sup>9, 12</sup>

$$F[n, m] = \sum_{k=0}^{N-1} f[n-k] w^{km} \quad w = e^{j2\pi/N} \quad (39)$$

where  $N$  is a given integer (see also Eq. (20)). For a fixed  $n$ ,  $F[n, m]$  is the DFS in the variable  $k$  of the segment  $f[n-k]$  of  $f[n]$  shown in Figure 5.

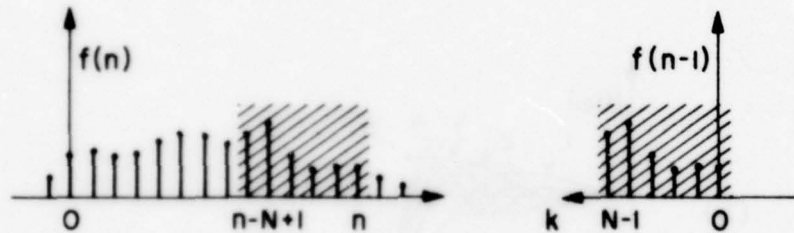


Fig. 5.

Inversion Formula: For any  $n$ ,

$$f[n] = \frac{1}{N} \sum_{m=0}^{N-1} F[n, m] \quad (40)$$

Proof. From the inversion formula for DFS it follows that

$$f[n-k] = \frac{1}{N} \sum_{m=0}^{N-1} F[n, m] w^{-km} \quad (41)$$

and Eq. (40) results with  $k=0$ .

Filtering: Suppose that  $f[n]$  is the input to a non-recursive filter of length  $N$ . The resulting output is given by<sup>13</sup>

$$g[n] = \sum_{k=0}^{N-1} f[n-k] h[k] \quad (42)$$

We maintain that<sup>14</sup>

$$g[n] = \sum_{m=0}^{N-1} F[n, m] H(w^m) \quad (43)$$

where

$$H(z) = \sum_{k=0}^{N-1} h[k] z^{-k} \quad (44)$$

is the system function.

Proof. Inserting Eq. (41) into Eq. (42) and changing the summation order, we obtain Equation (43).

Running Discrete Low-Pass Filters: If

$$H(w^m) = \begin{cases} 1 & |m| \leq M \\ 0 & |m| \geq M \end{cases} \quad (45)$$

then Eq. (43) yields

$$g[n] = \frac{1}{N} \sum_{m=-M}^M F[n, m] = \frac{1}{N} F[n, 0] + \frac{2}{N} \sum_{m=1}^M \operatorname{Re} F[n, m] \quad (46)$$

From the above and Eq. (41) it follows that

$$g[n] = \frac{1}{N} \sum_{m=-M}^M \sum_{k=0}^{N-1} f[n-k] w^{km} = \frac{1}{N} \sum_{k=0}^{N-1} f[n-k] \sum_{m=-M}^M w^{km}$$

Hence, the delta response of the running low-pass filter equals

$$h[n] = \frac{1}{N} \sum_{m=-M}^M w^{nm} = \frac{\sin \left[ \left( M + \frac{1}{2} \right) \frac{2\pi n}{N} \right]}{N \sin \frac{\pi n}{N}} \quad (47)$$

for  $n$  between 0 and  $N-1$  and it equals zero otherwise.

The running discrete Fourier filter is an extension of Eq. (47) and it is obtained by truncating the sum in Eq. (43) (see also Eq. (21)):

$$g[n] = \frac{1}{N} \sum_{m=-M}^M F[n, m] H(w^m) \quad (48)$$

The design of this filter involves the selection of the integer  $M$ , the determination of the weights  $H(w^m)$  and the computation of the sequence  $F[n, m]$ . We start with the last objective.

#### The DFS Analyzer

The number of multiplications needed to compute the running DFS  $F[n, m]$  from Eq. (41) equals  $N-1$ . This number can be reduced to a single multiplication if  $F[n, m]$  is computed recursively. Indeed, from Eq. (41) it follows that

$$F[n, m] - w^m F[n-1, m] = f[n] - f[n-N] \quad (49)$$

Hence,  $F[n, m]$  can be determined as the output of the first order system<sup>14</sup> shown in Figure 6(a).

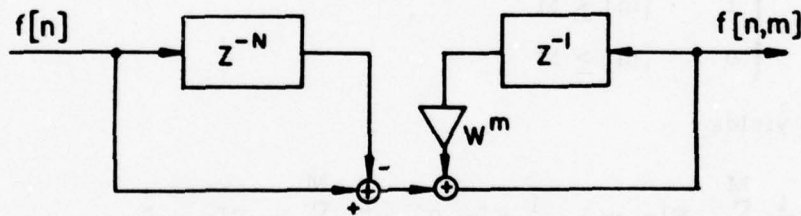


Fig. 6(a).

This system has a pole  $w^m$  on the unit circle. Furthermore, it requires a shift register or some other method for generating the delayed input  $f[n-N]$ . We discuss below a modification of Eq. (49) that avoids these difficulties.

The output  $F_\alpha[n, m]$  of the system of Fig. 6(b) satisfies the first order recursion

$$F_\alpha[n, m] - \alpha w^m F_\alpha[n, m] = f[n] \quad (50)$$

Solving, we obtain

$$F_\alpha[n, m] = \sum_{i=0}^{\infty} \alpha^i w^{im} f[n-i] \quad (51)$$

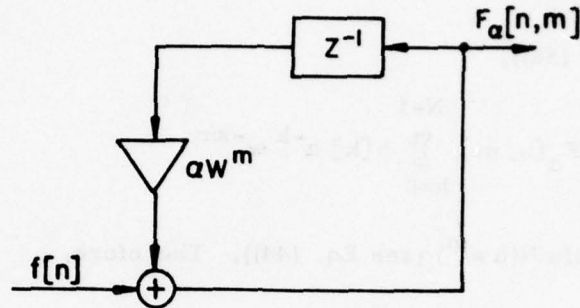


Fig. 6(b).

With  $i = k + rN$ , the above yields

$$F_{\alpha}[n, m] = \sum_{k=0}^{N-1} w^{km} \sum_{r=-\infty}^{\infty} \alpha^{k+rN} f[n-k-rN] \quad (52)$$

This shows that  $F_{\alpha}[n, m]$  is the DFS in the variable  $k$  of the last sum. Hence (inversion formula)

$$\sum_{r=-\infty}^{\infty} \alpha^{k+rN} f[n-k-rN] = \frac{1}{N} \sum_{m=0}^{N-1} F_{\alpha}[n, m] w^{-km} \quad (53)$$

We now assume that

$$\alpha^N \ll 1 \quad (54)$$

With this assumption, Eqs. (52) and (53) yield

$$F_{\alpha}[n, m] = \sum_{k=0}^{N-1} \alpha^k f[n-k] w^{km} \quad (55)$$

$$\alpha^k f[n-k] = \frac{1}{N} \sum_{m=0}^{N-1} F_{\alpha}[n, m] w^{-km} \quad (56)$$

We shall now express the output  $g[n]$  of a non-recursive filter of length  $N$  in terms of  $F_{\alpha}[n, m]$ . From Eq. (42) it follows that

$$g[n] = \sum_{k=0}^{N-1} \alpha^k f[n-k] \alpha^{-k} h[k]$$

Hence (see Eq. (56)),

$$g[n] = \frac{1}{N} \sum_{m=0}^{N-1} F_{\alpha}[n, m] \sum_{k=0}^{N-1} h[k] \alpha^{-k} w^{-km}$$

But the last sum equals  $H(\alpha w^m)$  (see Eq. (44)). Therefore,

$$g[n] = \frac{1}{N} \sum_{m=0}^{N-1} F_{\alpha}[n, m] H(\alpha w^m) \quad (57)$$

This is the modified form of Eq. (43) and it holds under the constraint Equation (54). The resulting filter preserves the frequency domain properties of the weights  $H(\alpha w^m)$  if  $\alpha$  is close to one. This condition is compatible with the constraint Eq. (54) if  $N$  is sufficiently large.

Truncating Eq. (61), we obtain

$$g[n] = \frac{1}{N} \sum_{m=-M}^M F_{\alpha}[n, m] H(\alpha w^m) \quad (58)$$

Equations (48) and (58) specify the running frequency domain filter. Its implementation is shown in Figure 7. It consists of  $2M+1$  analyzers as in Fig. 6, followed by  $2M+1$  weights  $H(w^m)$  or  $H(\alpha w^m)$  or  $H(\alpha w^m)$ , and an adder.

In the notations of Section A,  $g[n]$  equals the estimate  $\hat{y}[n]$  of a signal  $y[n]$  and  $f[n]$  equals the data sequence  $x[n]$ . We give next the recursive determination of the unknown weights limiting the discussion to the sum in Equation (48). The results are similar for Equation (58).

#### The Frequency Domain Wiener Filter

We denote  $B[n]$  a vector whose  $2M+1$  components  $b_m[n]$  are the adaptively controlled values of the weights  $H(w^m)$ . We wish to determine these components such that, if the sum

$$\hat{y}[n] = \sum_{m=-M}^M b_m[n] F[n, m] \quad (59)$$

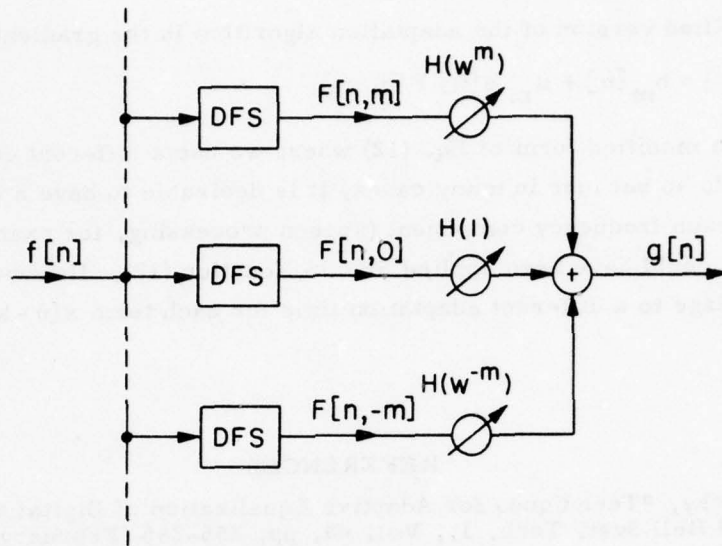


Fig. 7.

is used to estimate a signal  $y[n]$ , then the MS value of the resulting error

$$e[n] = y[n] - \hat{y}[n]$$

is minimum. In Eq. (53),  $F[n, m]$  is the running DFS of the data  $x[n]$ .

If the processes  $x[n]$  and  $y[n]$  are jointly stationary, with known second order moments, then the optimum vector  $B[n]$  is independent of  $n$  and it is given by (orthogonality principle)<sup>3</sup>

$$B = R^{-1} \Gamma \quad (60)$$

as in Equation (6). In the above,  $R$  is a matrix with elements

$$E\{F[n, m] F^*[n, m]\} \quad (61)$$

and  $\Gamma$  is a vector with elements

$$E\{F[n, m] y[n]\} \quad (62)$$

As we show in the section, these quantities can be expressed in terms of the moments of  $x[n]$  and  $y[n]$ . If these moments are not known, then the vector  $B[n]$  can be determined adaptively as in Equation (11).

### The Frequency Domain Widrow Filter

A simplified version of the adaptation algorithm is the gradient seeking recursion

$$b_m[n+1] = b_m[n] + \mu_m e[n] F[n, m] \quad (63)$$

This is a modified form of Eq. (12) where we use a different constant  $\mu_m$  for each  $m$ . We do so because in many cases, it is desirable to have a different adaptation time for each frequency component (speech processing, for example).<sup>15</sup> This generalization could have been applied also in Equation (12). However, there is no obvious advantage to a different adaptation time for each term  $x[n-k]$  of the sum in Equation (1).

A. Papoulis

### REFERENCES

1. R. W. Lucky, "Techniques for Adaptive Equalization of Digital Communication Systems," Bell Syst. Tech. J., Vol. 43, pp. 255-286 (February 1966).
2. J. G. Proakis and J. H. Miller, "An Adaptive Receiver for Digital Signaling through Channels with Intersymbol Interference," IEEE Trans. Inform. Theory, Vol. IT-15, pp. 484-497 (July 1969).
3. A. Papoulis, "Probability, Random Variables, and Stochastic Processes," New York: McGraw-Hill (1965).
4. B. Widrow and M. E. Hoff, "Adaptive Switching Circuits," 1960 WESCON Conf. Rev. Pt. 4, pp. 96-140.
5. B. Widrow et al, "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter," Proc. IEEE, Vol. 64, No. 8, pp. 1151-1162 (1976).
6. L. J. Griffiths, "Rapid Measurement of Digital Instantaneous Frequency," IEEE Trans. on Acoustics, Vol. ASSP-23, No. 2, pp. 207-222 (1975).
7. L. D. Davisson, "Steady-State Error in Adaptive Mean-Square Minimization," IEEE Trans. Inform. Theory, Vol. IT-19, pp. 382-385 (June 1970).
8. Jac-Kyoon Kim and L. D. Davisson, "Adaptive Linear Estimation for Stationary M-Dependent Processes," IEEE Trans. Inform. Theory, Vol. IT-21, No. 1, pp. 23-31 (January 1975).
9. A. Papoulis, "Signal Analysis," New York: McGraw-Hill (1977).
10. A. Papoulis, "The Fourier Integral and its Applications," New York: McGraw-Hill, (1962).
11. A. Papoulis, "Two-to-One Rule in Data Smoothing," IEEE Trans. Inform. Theory, Vol. IT-23 (September 1977).
12. J. B. Allen and L. R. Rabiner, "A Unified Theory of Short-Time Spectrum Analysis and Synthesis," Proc. IEEE (November 1977).
13. A. V. Oppenheim and R. W. Shafer, "Digital Signal Processing," Englewood Cliffs, N.J.: Prentice Hall, Inc. (1975).
14. L. R. Rabiner and R. W. Shafer, "Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques," IEEE Trans. on Audio and Electroacoustics, Vol. AU-19, No. 3, pp. 200-207 (September 1971).
15. J. B. Allen, D. A. Berkley and J. Blauert, "Multimicrophone Signal Processing Technique to Remove Room Reverberations from Speech Signals," J. Acoust. Soc. Am., Vol. 62, No. 4, pp. 912-915 (1977).

# BOUND ON THE PROBABILITY OF MISCLASSIFICATION OF TWO-DIMENSIONAL IMAGES

L. Kurz and P. Legakis

Recently the authors introduced an efficient statistical approach to image classification.<sup>1</sup> As a measure of performance of the procedure, asymptotic analysis and simulations were used. In this report, a bound on the probability of image misclassification based on the two-dimensional Chebyshev inequality<sup>2</sup> is introduced. The parameters in the inequality are obtained from appropriate simulation studies.

## A. The Bivariate Chebyshev System

Marshall and Olkin<sup>2</sup> have shown that if  $x$  and  $y$  are two random variables with known first and second order moments, and if  $R(x, y)$  is a non-negative risk function, then  $E[R(x, y)]$  can be bounded in terms of the first and second order moments. This bivariate Chebyshev system is useful if one wishes to establish a lower bound on the probability of correct classification. To relate the Chebyshev system to the problem at hand with the help of Fig. 1, let

$R_H$  Prob  $[(U, S) \in \text{Region H given that H is true}]$

$R_K$  Prob  $[(U, S) \in \text{Region K given that K is true}]$

$S$  The signal space

$U$  The statistic space

then

$$\text{Prob [correct classification]} \geq \frac{1}{2} (R_H + R_K) \quad (1)$$

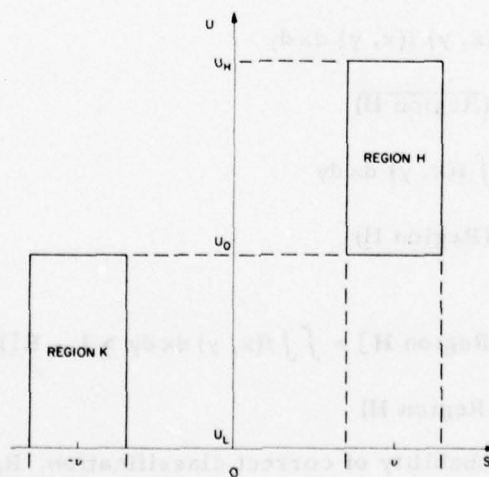


Fig. 1. Decision space for Chebyshev bound.

### Determination of $R_H$

For convenience let the variables  $x$  and  $y$  be defined as

$$x = S - E[S] \quad \text{and} \quad Y = \frac{v_o + \delta}{2} \quad (2)$$

so that

$$E[x] = \mu_x = 0; \quad E[y] = \mu_y = m_{U/H} - \frac{v_o + \delta}{2} \quad (3)$$

$$\text{Var}[x] = \sigma_x^2 = \sigma_s^2; \quad \text{Var}[y] = \sigma_y^2 = \sigma_{U/H}^2; \quad \sigma_{xy} = \sigma_{US}$$

and the region of interest is centered around the origin. All expectations are taken under  $H$  but subscripts are omitted for notational simplicity. Following Marshall and Olkin,  $R(x, y)$  must be quadratic, possibly with linear terms, so that

$$R(x, y) = ax^2 + by^2 + cxy + dx + ey \quad (4)$$

subject to

- a.  $R(x, y) \geq 1$ , for all  $x, y$  exterior to Region  $H$
  - b.  $R(x, y) \geq 0$ , otherwise
- (5)

Then

$$\begin{aligned} E[R(x, y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R(x, y) f(x, y) dx dy \\ &\geq \int \int R(x, y) f(x, y) dx dy \\ &\quad (x, y) \in (\overline{\text{Region } H}) \\ &\geq 1 - \int \int f(x, y) dx dy \\ &\quad (x, y) \in (\text{Region } H) \end{aligned} \quad (6)$$

and

$$\begin{aligned} R_H = \text{Prob}[(x, y) \in \text{Region } H] &= \int \int f(x, y) dx dy \geq 1 - E[R(x, y)] \\ &\quad (x, y) \in (\text{Region } H) \end{aligned} \quad (7)$$

To maximize the probability of correct classification,  $R_H$  must be made as large as possible subject to constraints in Equation (5). This is equivalent to maximizing

$E[R(x, y)]$  subject to the same constraints. The overall procedure is further simplified if normalization with respect to  $\sigma_x^2$  and  $\sigma_y^2$  is used. Let

$$x_o = \frac{x}{\xi_1 \sigma_x} \quad \text{and} \quad y_o = \frac{y}{\xi_2 \sigma_y} \quad (8)$$

so that region H is centered around the origin and is bounded by  $|x_o| \leq 1$  and  $|y_o| \leq 1$ . The coefficients of  $R(x_o, y_o)$  are chosen to provide a minimum for  $E[R(x_o, y_o)]$ . Coefficients d and e have been eliminated via normalization of first order moments. Thus,

$$R(x_o, y_o) = ax_o^2 + by_o^2 + cx_o y_o$$

The coefficients of Eq. (8) must be chosen to satisfy the constraints Equation (5).

For constraint a

Let

$$x_o = 0$$

Then

$$R(0, y_o) = by_o^2 \geq \begin{cases} 0 & |y_o| \leq 1 \\ 1 & \text{Otherwise} \end{cases}$$

$$\rightarrow b \geq 1$$

Let

$$y_o = 0$$

Then

$$R(x_o, 0) = ax_o^2 \geq \begin{cases} 0 & |x_o| \leq 1 \\ 1 & \text{Otherwise} \end{cases}$$

$$\rightarrow a \geq 1$$

For constraint b

Let

$$x_o = \pm 1$$

Then

$$R(\pm 1, y_o) = a + by_o^2 \pm cy_o$$

and

$$\frac{\partial R(\pm 1, y_o)}{\partial y_o} = 0 \rightarrow 2by_o = \mp c$$

Let

$$y_o = -\frac{c}{2b}, \text{ then } R(\pm 1, -\frac{c}{2b}) = a - \frac{c^2}{4b}$$

or

$$R(\pm 1, -\frac{c}{2b}) = 1 \rightarrow c^2 = 4b(a - 1)$$

Similarly, for

$$y_o = \pm 1, c^2 = 4a(b - 1)$$

and hence

$a = b$ . Letting

$$a = \frac{1}{1-k^2}, |k| < 1,$$

Eq. (7) can be written as

$$R(x_o, y_o) = \frac{1}{1-k^2} (x_o^2 + y_o^2 - 2kx_o y_o)$$

and the parameter  $k$  is chosen to minimize  $E[R(x, y)]$

$$R(x, y) = \frac{1}{1-k^2} \left( \left( \frac{x}{\xi_1 \sigma_x} \right)^2 + \left( \frac{y}{\xi_2 \sigma_y} \right)^2 - 2k \frac{x}{\xi_1 \sigma_x} \frac{y}{\xi_2 \sigma_y} \right)$$

$$\begin{aligned} E[R(x, y)] &= \frac{1}{1-k^2} \left\{ \frac{E[x^2]}{\xi_1^2 \sigma_x^2} + \frac{E[y^2]}{\xi_2^2 \sigma_y^2} - 2k \frac{E[xy]}{\xi_1 \sigma_x \xi_2 \sigma_y} \right\} \\ &= \frac{1}{1-k^2} \left\{ \frac{\frac{x^2}{\xi_1^2 \sigma_x^2} + \mu_x^2}{\xi_1^2 \sigma_x^2} + \frac{\frac{y^2}{\xi_2^2 \sigma_y^2} + \mu_y^2}{\xi_2^2 \sigma_y^2} - \frac{2k}{\xi_1 \xi_2} \left( \rho + \frac{\mu_x \mu_y}{\sigma_x \sigma_y} \right) \right\} \end{aligned}$$

where

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Letting

$$A_1 = \frac{1}{\xi_1^2} \left( 1 + \left( \frac{\mu_x}{\sigma_x} \right)^2 \right) + \frac{1}{\xi_2^2} \left( 1 + \left( \frac{\mu_y}{\sigma_y} \right)^2 \right)$$

$$A_2 = \frac{1}{\xi_1 \xi_2} \left( \rho + \frac{\mu_x \mu_y}{\sigma_x \sigma_y} \right)$$

Equation ( ) reduces to

$$E[R(x, y)] = \frac{1}{1-k^2} [A_1 - 2kA_2]$$

and

$$\frac{\partial E[R(x, y)]}{\partial k} = 0$$

leads to

$$A_2 k^2 - A_1 k + A_2 = 0$$

$$k_{1,2} = \frac{A_1}{2A_2} \pm \frac{1}{2} \sqrt{\left( \frac{A_1}{A_2} \right)^2 - 4}$$

Since  $|k| \leq 1$ , the negative sign is chosen and also, since  $k$  must be real,  $\frac{A_1}{A_2} \geq 2$ , so that

$$[E[R(x, y)]]_{\min} = \frac{A_2 \sqrt{\left( \frac{A_1}{A_2} \right)^2 - 4}}{1 - \frac{1}{4} \left[ \frac{A_1}{A_2} - \sqrt{\left( \frac{A_1}{A_2} \right)^2 - 4} \right]^2}$$

Thus

$$R_H = 1 - [E[R(x, y)]]_{\min} = 1 - \frac{4A_2 \sqrt{A_1^2 - 4A_2^2}}{4A_2^2 - [A_1^2 - \sqrt{A_1^2 - 4A_2^2}]^2}$$

Since Eq. ( ) was derived without assumptions on  $U$  and  $S$ , the expression for  $R_K$  is of a similar form.

$$R_K = 1 - \frac{4B_2^2 \sqrt{B_1^2 - 4B_2^2}}{4B_2^2 - [B_1^2 - \sqrt{B_1^2 - 4B_2^2}]^2}$$

where

$$B_1 = \frac{1}{\xi_{1/K}} \left[ 1 + \left( \frac{\mu_{x/K}}{\sigma_{x/K}} \right)^2 \right] + \frac{1}{\xi_{2/K}} \left[ 1 + \left( \frac{\mu_{y/K}}{\sigma_{y/K}} \right)^2 \right]$$

$$B_2 = \frac{1}{\xi_{1/K} \xi_{2/K}} \left( \rho_K + \frac{\mu_{x/K} \mu_{y/K}}{\sigma_{x/K} \sigma_{y/K}} \right)$$

Since

$$\text{Prob} [\text{correct classification}] \geq \frac{1}{2} (R_H + R_K)$$

the probability of misclassification is bounded by

$$P_e \leq 1 - \frac{1}{2} (R_H + R_K) = \frac{4A_2^2 \sqrt{A_1^2 - 4A_2^2}}{4A_2^2 - [A_1^2 - \sqrt{A_1^2 - 4A_2^2}]^2} + \frac{4B_2^2 \sqrt{B_1^2 - 4B_2^2}}{4B_2^2 - [B_1^2 - \sqrt{B_1^2 - 4B_2^2}]^2}$$

In order for  $P_e$  to be specified in terms of first and second order moments, a functional relationship between  $U$  and  $S$  must be found. Since it is analytically very difficult to find one, the simulation route was followed. Simulation under various signal conditions led to the functional forms of Figures 2 through 5. So, for low and moderate signal-to-noise ratios,  $U$  is related to  $S$  as follows:

$$\text{Under the hypothesis: } U = k_1 |S| + k_2$$

$$\text{Under the alternative: } U = -k_1 |S| + k_2$$

and

$$E_H[U] = \mu_{U/H} = k_1 \mu_{S/H} + k_2$$

$$\text{Var}_H[U] = \sigma_{U/H}^2 = k_1^2 \sigma_{S/H}^2$$

$$\text{Cov}_H[U, S] = k_1 \sigma_{S/H}^2$$

$$\rho_H = \frac{k_1 \sigma_{S/H}^2}{k_1 \sigma_{S/H} \sigma_{S/H}} = 1 = \rho$$

$$\xi_{1/H} = \frac{\Delta}{\sigma_{S/H}} = \frac{\Delta}{\sigma_{S/K}} = \xi_{1/K}$$

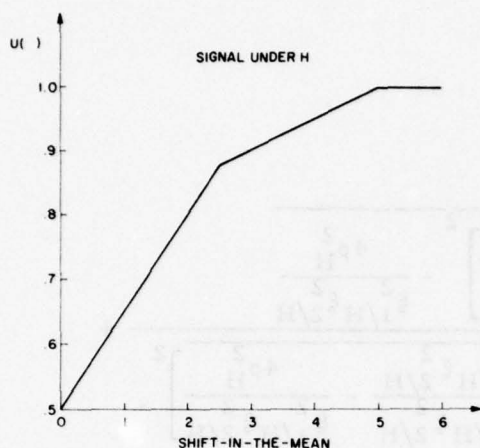


Fig. 2. Three-sample Mann-Whitney statistic versus shift-in-the-mean. Signal under H.

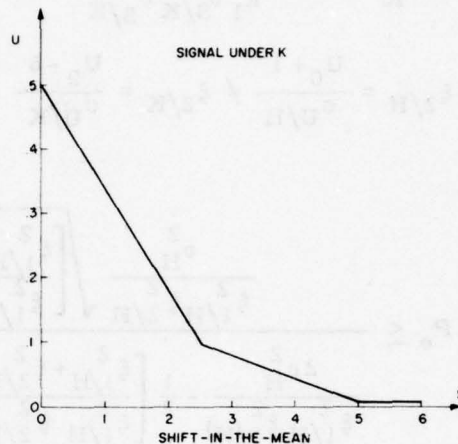


Fig. 3. Three-sample quantile Mann-Whitney statistic versus shift-in-the-mean. Signal under K.

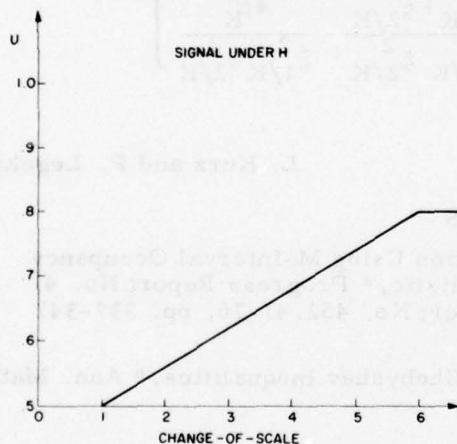


Fig. 4. Three-sample quantile Mann-Whitney statistic versus change-of-scale. Signal under H.

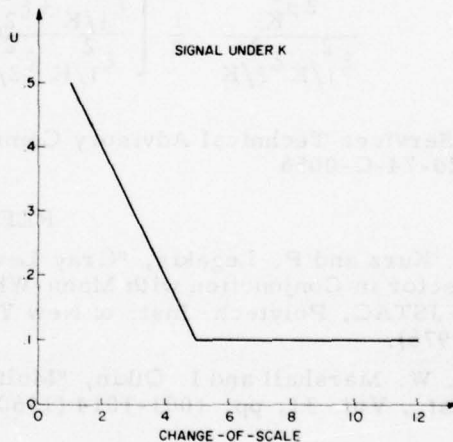


Fig. 5. Three-sample quantile Mann-Whitney statistic versus change-of-scale. Signal under K.

$$E_K[U] = \mu_{U/K} = -k_1 \mu_{S/H} + k_2$$

$$\text{Var}_K[U] = \sigma_{U/K}^2 = k_1^2 \sigma_{S/K}^2$$

$$\text{Cov}_K[U, S] = \frac{-k_1 \sigma_{S/K}}{k_1 \sigma_{S/K} \sigma_{S/K}} = -1 = -\rho$$

$$\xi_{2/H} = \frac{U_0 + 1}{\sigma_{U/H}} \neq \xi_{2/K} = \frac{U_0 - \delta}{\sigma_{U/K}}$$

Thus,

$$P_e \leq \frac{\frac{\rho_H^2}{\xi_{1/H}^2 \xi_{2/H}^2} \sqrt{\left[ \frac{\xi_{1/H}^2 + \xi_{2/H}^2}{\xi_{1/H}^2 \xi_{2/H}^2} \right]^2} - \frac{4\rho_H^2}{\xi_{1/H}^2 \xi_{2/H}^2}}{\frac{2\rho_H^2}{\xi_{1/H}^2 \xi_{1/H}^2} - \frac{1}{2} \left[ \frac{\xi_{1/H}^2 + \xi_{2/H}^2}{\xi_{1/H}^2 \xi_{2/H}^2} - \sqrt{\frac{\xi_{1/H}^2 + \xi_{2/H}^2}{\xi_{1/H}^2 \xi_{2/H}^2}} - \frac{4\rho_H^2}{\xi_{1/H}^2 \xi_{2/H}^2} \right]^2} +$$

$$+ \frac{\frac{\rho_K^2}{\xi_{1/K}^2 \xi_{2/K}^2} \sqrt{\left[ \frac{\xi_{1/K}^2 + \xi_{2/K}^2}{\xi_{1/K}^2 \xi_{2/K}^2} \right]^2} - \frac{4\rho_K^2}{\xi_{1/K}^2 \xi_{2/K}^2}}{\frac{2\rho_K^2}{\xi_{1/K}^2 \xi_{2/K}^2} - \frac{1}{2} \left[ \frac{\xi_{1/K}^2 + \xi_{2/K}^2}{\xi_{1/K}^2 \xi_{2/K}^2} - \sqrt{\frac{\xi_{1/K}^2 + \xi_{2/K}^2}{\xi_{1/K}^2 \xi_{2/K}^2}} - \frac{4\rho_K^2}{\xi_{1/K}^2 \xi_{2/K}^2} \right]^2}$$

Joint Services Technical Advisory Committee  
F44620-74-C-0056

L. Kurz and P. Legakis

#### REFERENCES

1. L. Kurz and P. Legakis, "Gray Level Detection Using M-Interval Occupancy Vector in Conjunction with Mann-Whitney Statistic," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. 452.41-76, pp. 337-347 (1976).
2. A. W. Marshall and I. Olkin, "Multivariate Chebyshev Inequalities," Ann. Math. Stat., Vol. 31, pp. 1001-1014 (1960).

## QUADRATIC TESTS IN M-GRAY LEVEL DETECTION

L. Kurz and P. Legakis

The generalization of the edge detection problem from two to M levels under moderate and high signal-to-noise ratio conditions has been treated by the authors elsewhere.<sup>1</sup> If images are severely corrupted by noise, as is usually the case in sonar and radar data, the procedures suggested in the past do not perform satisfactorily. In this report, an m-interval quadratic test<sup>2</sup> is used to develop an efficient procedure of M-gray level detection of images severely corrupted by noise. A typical result of implementation of the procedure to simulated data supports the theoretical results presented below. Also, a procedure for empirical establishment of decision thresholds is outlined.

A. Quadratic Tests

When the generalized signal-to-noise ratio,  $\lambda$  between two hypotheses  $H_k$  and  $H_{k+1}$  defined as

$$\lambda = \int_{-\infty}^{\infty} (F(x - \mu_k) - F(x - \mu_{k+1})) dF(x - \mu_k) \quad (1)$$

is moderate or large, as is the case in most image processing under noisy conditions, the statistic introduced in Ref. 1 gives good results and is very easy to implement. However, when  $\lambda$  is small, it performs poorly and a different class of statistics is recommended. Such is the set of quadratic statistics which belong to the general class of nonlinear m-interval, M-ary detection statistics.<sup>2</sup> These tests are based on M-interval equiprobably partitioning of the probability space and are of the "matched filter" variety. The choice of nonlinear over linear statistics is based on the fact that the latter are suitable for stochastic ordering type problems only while the nonlinear statistics are good for both stochastic ordering and change-of-scale problems. This property makes them good candidates for gray level and texture edge detection in image processing under severe noise conditions. The quadratic statistics are formulated as follows:

Based on reference sample  $X_1, X_2, \dots, X_M$ , quantile vectors  $\bar{\alpha}_k = (\alpha_{k1}, \dots, \alpha_{kM})$ ,  $k = 1, 2, \dots, M$ , are selected. The selection is made in such a way that when  $H_k$  holds, the observation space is equiprobably partitioned. Thus, if  $\bar{n}_k$  is the mapping of an observation vector  $Y_n$  on the partitions defined by  $\bar{\alpha}_k$  and if  $Y_n \sim F(x - \mu_k)$ , then  $\bar{n}_k$  is an accurate m-interval representation of  $\bar{Y}_n$  and the partition  $\bar{\alpha}_k$  is such that

$$E[\bar{n}_k] = \left[ \frac{n}{m}, \frac{n}{m}, \dots, \frac{n}{m} \right] \quad (2)$$

In addition, since  $\sum_{j=1}^n n_j = n$ , the quadratic statistic

$$T_k = \bar{n}_k' A \bar{n}_k \quad (3)$$

shows most atypical behavior when  $Y_n \sim F(x - \mu_k)$  in the sense that it attains a minimum. Thus, if  $Y_n$  is mapped into every  $\bar{a}_k$ ,  $k = 1, 2, 3, \dots, M$ , and the statistics  $T_k$  are formed according to Eq. (3), then the classification of  $Y_n$  is done on the basis of

$$\min(T_1, T_2, \dots, T_M)$$

For an extensive treatment of the small signal behavior of the three-sample quadratic statistic based on quantile occupancy vectors, its asymptotic comparison with the three-sample Mann-Whitney statistic, Ref. 2, is recommended.

#### B. An Empirical Establishment of Classifier Thresholds and Comparison of Test Statistics

It was noted in Ref. 1 that if the classification of a test sample is done in parallel,  $M$  test statistics are generated simultaneously and a decision is made on the basis of  $\min(T_1, T_2, T_3, \dots, T_M)$ . Hence, there is no need of classifier thresholds. A sequential procedure is slightly different in the sense that in order to distinguish between two hypothesis,  $H_k$  and  $H_{k+1}$  a single statistic  $T_k$  is formed and a decision is made on the basis of whether  $T_k$  exceeds or is less than a certain threshold. The usual measure of comparison of statistics is through the concept of asymptotic relative efficiency (ARE) which is the ratio of efficacies of the statistic. The efficacy  $\mathcal{E}$  of a statistic  $T_k$  is given by

$$\mathcal{E} = \lim_{\substack{n \rightarrow \infty \\ H_{k+1} \rightarrow H_k}} \frac{\frac{\partial}{\partial \theta} E[T_k/H_{k+1}]}{n \text{Var}[T_k/H_k]} \bigg|_{\theta=0}$$

Where  $n$  is the sample size and  $\theta$  is the parameter that separates the hypothesis from the alternative. Notice that  $\mathcal{E}$  as defined above has meaning only as  $n \rightarrow \infty$  and  $\theta \rightarrow 0$ . The observation to be made is that in both, the establishment of classifier thresholds and the comparison of statistics through the concept of ARE, the results are asymptotically valid. In many applications one is interested in neither extremely large samples, nor in cases where the alternative approaches the hypothesis, but rather in moderate signal-to-noise ratios and finite sample sizes. This is particularly true in image processing where the signal-to-noise ratios are usually moderate to large and one would rather use the single quantile statistic rather than the quadratic statistic which takes at least five times the processing time of the linear single quantile statistic.

An empirical procedure is now suggested which allows arbitrary sample size, qualitative and quantitative comparison of statistics as well as the establishment of optimal thresholds. The outline of the procedure is as follows:

Given a binary rectangular ( $L_p \times L_q$ ) array  $W$  that contains the same number of zeros and ones, generate a new array  $W'$  where  $w'_{ij}$  takes on the value of a test statistic. The test statistic of the  $i$ -th row and  $j$ -th column of  $W'$  is based on a sample which is drawn from the hypothesis or the alternative depending on whether the  $w_{ij}$  of  $W$  is a zero or a one, respectively. Thus  $W'$  is an estimator of  $W$ . Let  $L \leftarrow (W' > \text{threshold})$ . Then,  $L$  is an ( $L_p \times L_q$ ) binary array consisting of elements  $w_{ij}$  such that

$$l_{ij} = \begin{cases} 1 & \text{if } w'_{ij} > \text{threshold} \\ 0 & \text{if } w'_{ij} < \text{threshold} \end{cases} \quad (4)$$

Thus, the average probability of correct classification is given by

$$\frac{1}{L_p L_q} \sum_{i=1}^{L_p} \sum_{j=1}^{L_q} (l_{ij} = w_{ij})$$

where

$$(l_{ij} = w_{ij}) = \begin{cases} 1 & \text{if } w_{ij} \text{ and } l_{ij} \text{ are equal} \\ 0 & \text{Otherwise} \end{cases}$$

and the average probability or misclassification or error.  $P'_e$  is given by

$$P'_e = 1 - \frac{1}{L_p L_q} \sum_{i=1}^{L_p} \sum_{j=1}^{L_q} (l_{ij} = w_{ij})$$

Once the  $W'$  array has been obtained for a given procedure, the threshold can be adjusted to give the optimal settings. This procedure was tried for the quadratic quantile tests and the results appear in Figures 1 and 2. The optimal threshold was obtained in less than 10 iterations. By selecting well known patterns as the array  $W$  one can make a qualitative comparison of two procedures by printing images of the reference pattern and using the optimal threshold for each procedure. This is done in Figures 3 through 5.

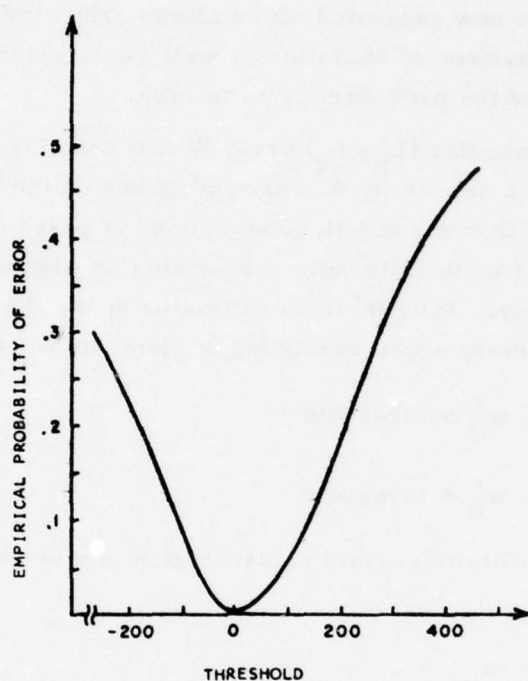


Fig. 1. Empirical prob. of error versus threshold.  
Three-sample quantile quadratic statistic.  
( $m=10$ ,  $n=100$ ,  $N(0,1)$ ,  $N(.5,1)$ ).

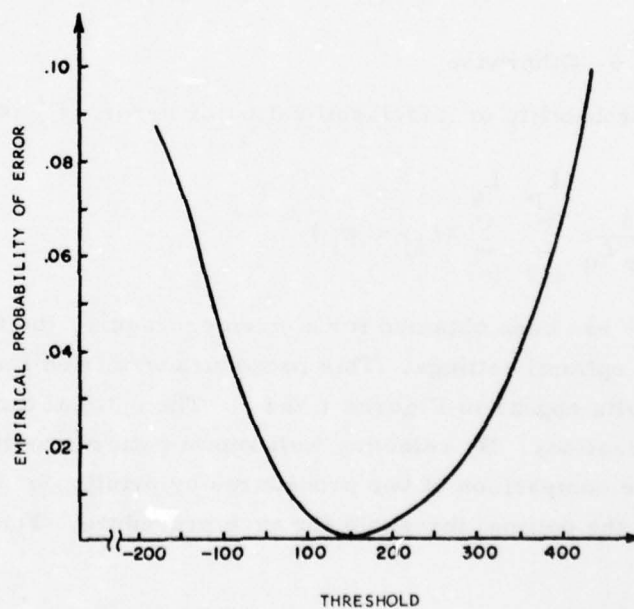


Fig. 2. Empirical prob. of error versus threshold.  
Three sample quantile quadratic statistic.  
( $m=10$ ,  $n=100$ ,  $N(0,1)$ ,  $N(0,.5)$ ).

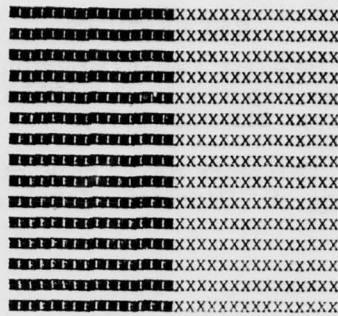


Fig. 3. Uncorrupted binary reference pattern.

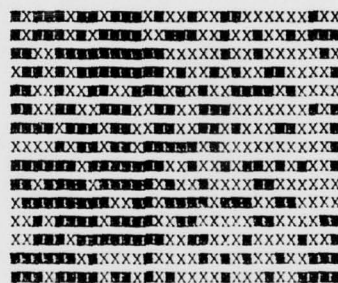


Fig. 4. Gray level detection based on single quantile statistics. Reference pattern corrupted with Gaussian  $N(0, 1)$  and  $N(.5, 1)$  noise.

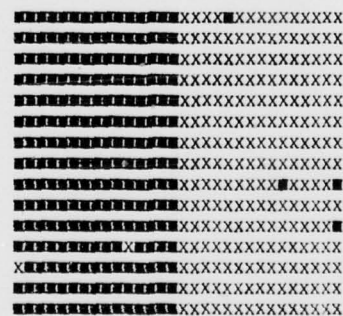


Fig. 5. Gray level detection based on quadratic quantile statistics. Binary levels of reference pattern corrupted with Gaussian  $N(0, 1)$  and  $N(.5, 1)$  noise.

## REFERENCES

1. P. Legakis and L. Kurz, "M-Gray Level Detection for Moderate and High Signal-to-Noise Ratios," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77, pp. 515-523 (1977).
2. L. Kurz, "Nonparametric Detectors Based on Partition Tests" in Nonparametric Methods in Communications, P. Papantoni-Kazakos and D. Kazakos, Marcel Dekker, 1977.

# ROBUSTIZED VECTOR FORM OF GLADYSHEV'S THEOREM WITH APPLICATION TO ESTIMATION OF PARAMETERS IN A LINEAR MODEL

I. Kadar and L. Kurz

In this report, the theory of robustized stochastic approximations of Gladyshev's form, presented by the authors in the scalar form in a recent report,<sup>1</sup> is extended to the vector case. The robustizing is accomplished along the lines suggested by Kersten and Kurz.<sup>2</sup> The resulting algorithm is then applied to the important problem of parameter estimation in a linear model.

## A. Vector Extension of Gladyshev's Theorem (SAMVLS)

Consider the linear model in its vector form by having multiple observations of  $y, y_k$  available. Then

$$y_k = X\beta_k + V_k, \quad k = 1, 2, \dots \quad (1)$$

The vector SAMVLS formulation (see Ref. 1 for the scalar formulation) for (1) becomes

$$\hat{\beta}_{k+1} = \hat{\beta}_k - (A_k/k) [X^T X \hat{\beta}_k - X^T y_k]$$

where,  $A_k$  is an  $(r \times r)$  diagonal adaptive gain matrix which satisfies the conditions of Theorem 1 of Reference 2.\*

$$Y(\hat{\beta}, \beta) = X^T(X\hat{\beta} - y) = X^T X(\hat{\beta} - \beta) + X^T V$$

and the associated regression function

$$M(\hat{\beta}, \beta) = X^T X(\hat{\beta} - \beta)$$

satisfies conditions (iv) to (vi), where it is clear with  $\alpha = \underline{0}$  that  $M(\hat{\beta}, \beta) = \alpha = \underline{0}$  has a unique root at  $\hat{\beta} = \beta$ . The B matrix  $B = X^T X = I_{r \times r}$  (by the definition of the problem) is a positive definite  $r \times r$  matrix s.t.  $\|\beta\| < \infty$  which is satisfied with  $X$  orthogonal s.t.  $X^T X = I$ . This makes  $P = I$  and  $A_k$  a diagonal matrix for each  $k$ .

The additive noise term

$$Z(\hat{\beta}) = X^T V$$

satisfies condition (ii) with  $EZ(\hat{\beta}) = \underline{0}$ . It must have uniformly bounded variance, as in the scalar case, and well-defined covariance matrix as  $\hat{\beta} \rightarrow \beta$ , a.s. Condition (viii),  $\sup E\|Z(\hat{\beta})\|^{2+2\epsilon} < \infty$  for some  $\epsilon > 0$  is satisfied with

$$\sup_{\hat{\beta}} [tr(X^T \Gamma X)]^{1+\epsilon} = \left[ \sum_{i=1}^r \sigma_{ii}^2 \right]^{1+\epsilon} < \infty;$$

\*Conditions mentioned hereafter refer to Theorem 1 of Reference 2.

$$\lim_{\hat{\beta} \rightarrow \beta} E[Z(\hat{\beta}) Z^T(\hat{\beta})] = \pi,$$

where  $\pi$  is a non-negative definite matrix, is satisfied with

$$\lim_{\hat{\beta} \rightarrow \beta} X^T \Gamma X = \sigma^2 [I]_{r \times r} = \pi.$$

Condition (viii) for the adaptive gain matrix is satisfied. Let  $a_1^{(k)} \geq a_2^{(k)} \geq \dots \geq a_r^{(k)} > 0$  be the eigenvalues of  $A_k$  with  $A_k$  diagonal. Similarly,  $b_1 \geq b_2 \geq \dots \geq b_r > 0$  are eigenvalues of  $B$ , which is diagonal and  $a_1 \geq a_2 \geq \dots \geq a_r > 0$  are eigenvalues of  $A$ , where

$$0 < a'_1 \leq \inf_{\hat{\beta}} \left\{ \sum_{i=1}^r [a_i^{(k)}]^2 \right\}^{1/2} \leq \sup_{\hat{\beta}} \left\{ \sum_{i=1}^r [a_i^{(k)}]^2 \right\}^{1/2} \leq a'' < \infty \text{ wpl},$$

for  $k$  large and  $\lim_{k \rightarrow \infty} a_r^{(k)}(\cdot) \geq a'_1 > 0$  wpl, where  $A$  is a constant matrix s.t.  $a'_1 \leq \left[ \sum_{i=1}^r a_i^2 \right]^{1/2} \leq a''$  since  $A$  is diagonal. Condition (ix)  $a'_1 b_r - \epsilon > 1/2$  is clearly satisfied with distinct nonzero eigenvalues of  $A$  and  $B$ , both diagonal.

### Theorem

Under assumptions (i) to (ix) stated in Theorem 1 of Ref. 2 and shown to be satisfied above, let  $r_1 > r_2 > \dots > r_r > 0$  be eigenvalues of  $AB$ ,  $B = X^T X = I_{r \times r}$ . Then  $k^{1/2}(\hat{\beta}_k - \beta)$  is asymptotically normal with mean zero and covariance matrix  $Q$  where  $Q$  is a diagonal matrix whose elements are  $a_{ii}^2 \sigma_{ii}^2 [2a_{ii} - 1]^{-1}$ . For proof see Reference 2.

### B. Robust Estimation of the Parameters in a Linear Model

We apply the theory established in the previous section to the estimation of the parameters of a reparameterized, replicated, full-rank, two-way layout with  $K$  observations per cell fixed effects experimental design model (i.e., a linear model).

In matrix notation, the model equation becomes,

$$y_k = X\beta_k + V_k, \quad k = 1, 2, \dots \quad (2)$$

where  $X$  could be, for instance, a  $(4 \times 3)$  orthogonal design matrix

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}, \quad X^T X = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

$\beta_k$  is a  $(3 \times 1)$  vector of parameters and  $V_k$  is a  $(4 \times 1)$  additive symmetrically distributed noise vector, i.i.d for each  $k$ .

To robustize Eq. (2), one introduces batch pre-processing and a nonparametric statistic (in a manner similar to Ref. 1 for the scalar case) applied to each ( $r=3$ ) component of

$$Y(\hat{\beta}_{k+i}, \beta) = X^T [X(\hat{\beta}_{k+i} - \beta) + V_{k+i}], \quad i = 1, 2, \dots, m,$$

say

$$S_\ell^m(\cdot) = \sum_{i>j}^m \text{sgn}(Z_i + Z_j) + \sum_{i=1}^m \text{sgn}(Z_i), \quad \ell = 1, 2, 3$$

where  $S_\ell^m(\cdot)$  in this case is a  $3 \times 1$  vector operator, and the robust vector operator, and the robust vector SAMVLS is of the form:

$$\hat{\beta}_{k+1} = \hat{\beta}_k - \frac{1}{k} A_k S^m Y(\hat{\beta}_k, \beta) \quad (3)$$

$S_\ell^m(\cdot)$  for each  $\ell = 1, 2, 3$  is a symmetric version of WSRNS. The use of  $S^m(\cdot)$  only requires that the class of noise distributions be symmetric and that  $S^m(Y)$  satisfy conditions (iv) to (vi) of Theorem 1 of Reference 2.

To obtain the components of the optimum gain matrix, one must find the appropriate component of the robustized regression function and evaluate the derivative at  $\hat{\beta} - \beta$  in the same manner as the scalar case. To accomplish this, define for each  $r$  ( $r=3$ ) components

$$U_\ell^m = \sum_{i=1}^m \sum_{j=1}^{i-1} U(Z_i + Z_j) + \sum_{i=1}^m U(Z_i), \quad \ell = 1, 2, 3$$

where,

$$\begin{bmatrix} S_1^m(\cdot) \\ S_2^m(\cdot) \\ S_3^m(\cdot) \end{bmatrix} = \begin{bmatrix} 2U_1^m(\cdot) - m(m+1)/2 \\ 2U_2^m(\cdot) - m(m+1)/2 \\ 2U_3^m(\cdot) - m(m+1)/2 \end{bmatrix}$$

Then it can be shown by adapting results from the scalar case to the vector case, that for each  $r=3$  component of the regression function with  $X^T X = 4[I]_{3 \times 3}$

$$EU^m = \begin{bmatrix} \sum_{i=1}^m \sum_{j=1}^{i-1} \int_{-4(\beta_1 - \hat{\beta}_{1k})}^{\infty} f_1(x) dx + \sum_{i=1}^m \int_{-4(\beta_1 - \hat{\beta}_{1k})}^{\infty} g_1(x) dx \\ \sum_{i=1}^m \sum_{j=1}^{i-1} \int_{-4(\beta_2 - \hat{\beta}_{2k})}^{\infty} f_2(x) dx + \sum_{i=1}^m \int_{-4(\beta_2 - \hat{\beta}_{2k})}^{\infty} g_2(x) dx \\ \sum_{i=1}^m \sum_{j=1}^{i-1} \int_{-4(\beta_3 - \hat{\beta}_{3k})}^{\infty} f_3(x) dx + \sum_{i=1}^m \int_{-4(\beta_3 - \hat{\beta}_{3k})}^{\infty} g_3(x) dx \end{bmatrix}$$

where,  $f_\ell(x)$ ,  $\ell = 1, 2, 3$  are the components of the pdf of  $X^T(V_1 + V_j)$  and  $g_\ell(x)$ ,  $\ell = 1, 2, 3$  are the components of the pdf of  $X^T V_i$ . The slope of the regression function is given by, for each of the  $r=3$  components,

$$B = \begin{bmatrix} \left[ 2 \sum_{i=1}^m \sum_{j=1}^{i-1} 4f_1(0) + 2 \sum_{i=1}^m 4g_1(0) \right] & 0 & 0 \\ 0 & \left[ 2 \sum_{i=1}^m \sum_{j=1}^{i-1} 4f_2(0) + 2 \sum_{i=1}^m 4g_2(0) \right] & 0 \\ 0 & 0 & \left[ 2 \sum_{i=1}^m \sum_{j=1}^{i-1} 4f_3(0) + 2 \sum_{i=1}^m 4g_3(0) \right] \end{bmatrix}$$

and the diagonal elements of B become

$$\alpha_{\ell\ell} = 8 \left[ m(m-1) f_{\ell\ell}(0) + m g_{\ell\ell}(0) \right], \quad \ell = 1, 2, 3$$

where  $f_{\ell\ell}(0)$  and  $g_{\ell\ell}(0)$  are the maximum values of the pdf's of  $r=3$  components of the vectors

$$\begin{bmatrix} 2(V_1 + V_2 + V_3 + V_4) \\ 2(V_1 + V_2 - V_3 - V_4) \\ 2(V_1 - V_2 + V_3 - V_4) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} (V_1 + V_2 + V_3 + V_4) \\ (V_1 + V_2 - V_3 - V_4) \\ (V_1 - V_2 + V_3 - V_4) \end{bmatrix}$$

respectively.

If the pdf of  $V_k$  is known for each  $k$ , then the elements of the optimum gain matrix, A, in terms of the  $\alpha_{\ell\ell}$ 's can be computed. However, the symmetric pdf of the i.i.d. vector sequence,  $V_k$ , may not be known. This means that we have to find an estimator of both  $f_{\ell\ell}(0)$  and  $g_{\ell\ell}(0)$ ,  $\ell = 1, 2, 3$  in order to find the elements of the

adaptive gain matrix. To this end, we represent the unknown class of cdf's by a vector extension of the generalized Gaussian noise

$$g^c(x) = \prod_{\ell=1}^3 g_{\ell}^c(x)$$

where  $g_{\ell}^c(x) = [c/2A(c)\Gamma(1/c)] \exp\{-|x_{\ell}|/A(c)\}^c$ ,  $\ell = 1, 2, 3$   $A(c) = [\sigma_{\ell}^2 \Gamma(1/c)\Gamma(3/c)]^{1/2}$  and the covariance of  $g^c(x)$  is diagonal with elements  $\sigma_{\ell\ell}^2$ ,  $\ell = 1, 2, 3$ . The range of values of  $g_{\ell}^c(0)$ ,  $\ell = 1, 2, 3$  between  $1 \leq c \leq 4$  (which represent a large class of CDF's) is only 2:1. This means that the diagonal elements of  $\alpha_{\ell\ell}$  can be approximated as

$$\alpha_{\ell\ell} \approx 8m^2 g_{\ell\ell}(0), \quad \ell = 1, 2, 3.$$

Now to obtain the diagonal elements of the adaptive gain matrix, a robust estimator of  $[8m^2 g_{\ell\ell}(0)]^{-1}$  is formed<sup>1,2</sup> which, in terms of the elements of  $A_k(\cdot)$  is given by

$$[A_k] = \frac{m+1}{8(k-1)} \sum_{j=1}^{k-1} \left[ Z_{j, [\frac{m}{2}]+1}^{\ell\ell} - Z_{j, [\frac{m}{2}]}^{\ell\ell} \right], \quad \ell = 1, 2, 3$$

where  $Z_{j, [\frac{m}{2}]+1}^{\ell\ell} \equiv [\frac{m}{2}]+1$ -th order statistic from the component of random samples  $y_{i+m(k-i)}^{\ell}$ ,  $i = 1, 2, m$ ,  $\ell = 1, 2, 3$  with pdf  $g(x) = \prod_{\ell=1}^3 g_{\ell}(x)$ , where  $[\frac{m}{2}]$  is defined to be the greatest integer less than or equal to  $\frac{m}{2}$ .

Reflecting back at this point to the requirements of Theorem 1 of Ref. 2, we see, from condition (vi) that  $P = [I]_{3 \times 3}$  with  $B = 4[I]_{3 \times 3}$ . The additive noise term  $Z(\hat{\beta}_k) = S^m Y(\cdot) - M(\cdot)$  has been shown to satisfy condition (vii) for the same functional form; and the adaptive diagonal gain matrix  $A_n(\cdot)$  with  $B = 4[I]_{3 \times 3}$  has also been shown to satisfy condition (viii) in Ref. 2 for a similar form of the batch-order statistic estimator of the elements of  $A_k(\cdot)$ . Thus, a flexible robustized algorithm for estimating the parameters of the linear model has been generated. Applications of this algorithm to problems of image processing, pattern recognition, radar and sonar data processing are the subject for future investigation.

Joint Services Technical Advisory Committee  
F44620-74-C-0056

I. Kadar and L. Kurz

Grumman Aerospace Corporation

## REFERENCES

1. I. Kadar and L. Kurz, "Robustized Scalar Form of Gladyshev's Theorem with Applications to Nonlinear Systems," Progress Report No. 42 to JSTAC, PINY, Report No. R-452.42-77, pp. 488-499 (1947).
2. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Information Sciences, Vol. 11, No. 2, pp. 121-140 (1976).

## MEETING PAPERS

Fourth European Geophysical Society Meeting, Munich, Germany, September 1977

S. H. Gross and H. Eun, "Relative Phase and Amplitude Response to Gravity Waves of Minor Species in the Thermosphere"

International Workshop on Optical Waveguide Theory, Reisenburg bei Grunzburg, W. Germany, September 1977

T. Tamir, "Microwave Network Approach to Planar Optical Components"

Seventh European Microwave Conference, Copenhagen, Denmark, September 1977

S. T. Peng and A. A. Oliner, "Leakage and Resonance Effects on Guiding Structures for Optical and Millimeter Waves"

IEEE COMPCON 77, Washington, D. C., September 1977

M. L. Shooman and A. Laemmel, "Statistical Theory of Computer Programs - Information Content and Complexity"

H. L. Bertoni, A. C. Green and L. B. Felsen, "Shadowing of an Inhomogeneous Plan Wave by an Edge"

Lawrence Symposium on Systems and Decision Sciences, Berkeley, California, October 1977

J. Hsuan and L. Shaw, "Optimal Control and Dynamic Repairman Assignment for a Linear Stochastic System"

Meeting of Operations Research Society, Atlanta, Georgia, November 7, 1977

C-L. Hsu, L. Shaw and S. G. Tyan, "Optimal Replacement Where Steps of Deterioration Have Multivariate Exponentially Distributed Durations"

Topical Meeting on Integrated and Guided Wave Optics, Salt Lake City, Utah, January 16-18, 1978

E-W. Hu, S. T. Peng and A. A. Oliner, "A Novel Leaky-Wave Strip Waveguide Directional Coupler"

Workshop on Large-Scale Systems, ORSA-TIMS Meeting, New York City, March 1978

R. Boorstyn, "Algorithms for Centralized Teleprocessing Network Design"

IEEE COMPCON 78, Washington, D.C., March 1978.

D. Baggi and M. L. Shooman, "An Automatic Driver for Pseudo-Exhaustive Software Testing"

9th Annual Pittsburgh Conference on Modeling and Simulation, Pittsburgh, Pennsylvania, April 1978

J. T. Doyle and L. Shaw, "Cyclically Updated Vehicle Control"

IEEE International Conference on Plasma Science, Monterey, California, May 15-17, 1978

S. P. Kuo and B. R. Cheo, "Harmonic Generation of the Electrostatic Ion Cyclotron Wave in a Uniform M-Plasma"

1978 International AP-S Symposium and URSI Spring Meeting, Washington, D. C., May 15-19, 1978

- J. P. Hsu, S. T. Peng and A. A. Oliner, "Scattering by Dielectric Step Discontinuities for Obliquely Incident Surface Waves"
- T. Ishihara and L. B. Felsen, "High Frequency Fields Excited by a Line Source Located on a Concave Cylindrical Impedance Surface"
- H. Steyskal, A. Hessel and J. Shmoys, "Limitations on Gain vs. Scan for a Dome Antenna"
- P. K. Bondyopadhyay, "Scattering by a Spherical Shell with an Arbitrarily Located Circular Aperture - A New Approach"
- A. Hessel, S. T. Peng and J. Shmoys, "Blazing of Rectangular Bar Transmission Gratings"

Conference on Coherent and Non-Linear Optics, Leningrad, USSR, June 1978

- T. Tamir and K. C. Chang, "Analysis and Design of Blazed Grating Couplers"

IFAC 7th World Congress, Helsinki, Finland, June 1978

- L. Shaw, D. Sarlat and Y. Thomas, "Synthesis of Nonlinear Controllers"

International Microwave Symposium, Ottawa, Canada, June 27-29, 1978

- L. B. Felson, "Review of Techniques for Propagation in Slab and Fiber Waveguides"
- P. K. Bondyopadhyay and A. Hessel, "Mutual Coupling Between Two Circular Waveguides Terminated in a Conducting Spherical Cavity"
- A. A. Oliner, S. T. Peng and J. P. Hsu, "New Propagation Effects for the Inverted Strip Dielectric Waveguide for Millimeter Waves"

International URSI XIX General Assembly, Helsinki, Finland, July 31-August 8, 1978

- S. W. Rosenthal, "Opening Remarks and Greetings - Biological Effects: State-of-the-Art-Review"
- A. A. Oliner and S. T. Peng, "Leaky Modes on Optical Strip Waveguides"
- A. C. Green, H. L. Bertoni and L. B. Felsen, "Properties of the Shadow Cast by a Half-Screen When Illuminated by a Gaussian Beam"
- K. C. Chang and T. Tamir, "Guiding and Scattering by Blazed Dielectric Gratings"
- L. B. Felsen, "High-Frequency Excitation of Concave Surfaces"
- A. A. Oliner, "Waveguides for Millimeter and Submillimeter Waves"
- A. Hessel, "Periodic Structure and GTD Method for Arrays on Concave Surfaces"

AFOSR Workshop on Communication Systems and Applications, Provincetown, Massachusetts, September 1978

R. Boorstyn, "Adaptive Routing in Networks"

#### JOURNAL ARTICLES AND CORRESPONDENCE

- E. Banks, S. Nakajima, L. C. Shapiro, O. Tilevitz, J. R. Alonzo and R. R. Chianelli, "Fibrous Apatite Grown on Modified Collagen," *Science*, Vol. 198, pp. 1164-1166 (1977).
- E. Banks, G. Torre and J. A. DeLuca, "Iron-57 Mossbauer Study of the Distribution of Divalent and Trivalent Ions in Potassium Transition Metal Fluorides Having Tetragonal Bronze Structure," *J. Solid State Chem.*, Vol. 22, pp. 95-100 (1977).
- S. Barone and P. Melman, "Asymptotic Solution of the Dirac Equation and the Inertial Mass of an Electron," *Phys. Rev. A.*, Vol. 19, No. 3, pp. 1115-1118 (Sept., 1978).
- H. L. Terton, A. C. Green and L. B. Felsen, "Shadowing of an Inhomogeneous Plane Wave by an Edge," *J. Opt. Soc. Am.* (1978).
- F. A. Cassara, "A Laboratory Experiment on Communication Electronics," *Int. J. Electr. Engr. Ed.*, Vol. 14, pp. 319-324 (September 1977).
- E. C. Cassedy and M. Jain, "A Theoretical Study of Injection Tuning of Optical Parametric Oscillators," submitted for publication *IEEE Trans. Quant. Electron.* (April 1978).
- K. K. Chan, L. B. Felsen, A. Hessel and J. Shmoys, "Creeping Waves on a Perfectly Conducting Cone," *IEEE Trans. Ant. Prop.*, Vol. AP-25, No. 5, pp. 661-670 (September 1977).
- K. K. Chan and L. B. Felsen, "The Pulsed Field Due to an Electric Dipole in the Presence of a Perfectly Conducting Wedge," *IEEE Trans. Ant. Prop.*, Vol. AP-25, pp. 420-423 (1977).
- K. K. Chan and L. B. Felsen, "Transient and Time-Harmonic Diffraction by a Semi-Infinite Cone," *IEEE Trans. Ant. Prop.*, Vol. AP-25, No. 6, pp. 802-806 (November 1977).
- S. Choudhary and L. B. Felsen, "Guided Modes in Graded Index Optical Fibers," *J. Opt. Soc. Am.*, Vol. 67, No. 9, pp. 1192-1196 (September 1977).
- S. Choudhary and L. B. Felsen, "Asymptotic Theory of Ducted Propagation," *J. Acous. Soc. Am.*, Vol. 63, pp. 661-666 (March 1978).
- R. S. Chu, J. A. Kong and T. Tamir, "Diffraction of Gaussian Beams by a Periodically Modulated Layer," *J. Opt. Soc. Am.*, Vol. 67, pp. 1555-1561 (November 1977).
- K. Chung and H. Eun, "Observation of Microwave Mixing in Air Glow Plasma," *IEEE Trans. on Plasma Sci.*, Vol. PS-5, pp. 41-48 (1977).
- R. F. Dwyer and L. Kurz, "Sequential Partition Detectors with Dependent Sampling," 1977 *IEEE Symp. on Infor. Theory*, p. 47 (October 1977).

- R. F. Dwyer and L. Kurz, "Sequential Partition Detectors," *Cybernetics* (May 1978).
- E. E. Kunhardt and B. R. Cheo, "Propagation of Non-Linear Waves Along a Magneto Plasma Column," *Phys. Fluids*, Vol. 20, No. 9, pp. 1499 (September 1977).
- E. E. Kunhardt and B. R. Cheo, "Experiments on Propagation of High Amplitude Surface Waves," accepted for publication in *Plasma Physics*.
- S. P. Kuo and B. R. Cheo, "Parametric Excitation of Coupled Plasma Waves," *Phys. Fluids*, Vol. 21, No. 10, pp. 1753 (1978).
- L. B. Felsen and C. Santana, "Ray Optical Calculation of Edge Diffraction in Unstable Resonators," to be published in *IEEE Transactions on Microwave Theory and Techniques*.
- L. B. Felsen and C. Santana, "Effects of Medium and Gain Inhomogeneities in Unstable Resonators," *Applied Optics*, Vol. 16, pp. 1058-1066 (1977).
- K. K. Chan and L. B. Felsen, "The Pulsed Field due to an Electric Dipole in the Presence of a Perfectly Conducting Wedge," *IEEE Trans. on Ant. Prop.*, Vol. AP-25, pp. 420-423 (1977).
- K. K. Chan, L. B. Felsen, A. Hessel and J. Shmoys, "Creeping Waves on a Perfectly Conducting Cone," *IEEE Trans. on Ant. Prop.*, pp. 661-670 (September 1977).
- K. K. Chan and L. B. Felsen, "Transient and Time-Harmonic Diffraction by a Semi-Infinite Cone," *IEEE Trans. on Ant. Prop.*, Vol. AP-25, No. 6, pp. 802-806 (November 1977).
- M. Greenblatt and E. Banks, " $\text{CdF}_2:\text{YbF}_3\text{ErE}_3$  - An Efficient Infrared to Visible Upconverting System," *J. Electrochem. Soc.*, Vol 124, p. 559 (1977).
- A. Hessel, L. Cheo and J. Shmoys, "On Simultaneous Blazing of Triangular Groove Diffraction Gratings," *J. Opt. Soc. Am.*, Vol. 67, No. 12 (December 1977).
- A. Hessel, H. S. Stalzer and J. Shmoys, *IEEE Trans. on Antennas and Prop.*, Vol. AP-26, No. 2 (march 1978).
- J. P. Hsu, S. T. Peng and A. A. Oliner, "Scattering by Dielectric Step Discontinuities for Obliquely Incident Surface Waves," *Digest of URSI Meeting*, p. 46, College Park, Maryland (May 1978).
- E. W. Hu, S. T. Peng and A. A. Oliner, "A Novel Leaky-Wave Strip Waveguide Directional Coupler," co-author, *Technical Digest, Topical Meeting on Integrated and Guided Wave Optics*, pp. WD2-1 to WD2-4, Salt Lake City, Utah (January 1978).
- H. M. Huang, S. P. Kuo and B. R. Cleo, "Harmonic Generation of the Electrostatic Ion Cyclotron Wave in a Uniform Magnetoplasma," *Bull. Am. Phys. Soc.*, Vol. 23, No. 7, p. 818 (September 1978).
- C. Hwa and L. M. Silber, "Ferromagnetic Resonance in Hexagonal Ferrites with Anisotropic g-Factors," *Physica*, Vol. 86-88 B, pp. 1239-1240 (1977).

- T. Ishihara, L. B. Felsen and A. Green, "High Frequency Fields Excited by a Line Source Located on a Perfectly Conducting Concave Cylindrical Surface," to be published in IEEE Trans. on Ant. and Prop.
- H. J. Juretschke, "Simple Derivation of the Maxwell Stress Tensor and Electrostrictive Effects in Crystals," A. J. Phys., Vol. 45, p. 277 (1977).
- P. R. Kersten and L. Kurz, "Improved Operation of m-Interval Detectors by Optimum Signal Selection," IEEE Trans. Inform. Theory, Vol. IT-24, No. 4, pp. 477-484 (July 1978)
- F. Kozin and T. S. Lee, "Almost Sure Asymptotic Likelihood Theory for Diffusion Processes," Vol. 14, pp. 527-537 (1977).
- H. Kudyan, "Interpretation of Electrostatic Energy Analyzer Data of a Flowing Plasma," Rev. Sci. Instrum., Vol. 49, No. 1, pp. 8-10 (January 1978).
- I. J. Kurland and H. L. Bertoni, "Birefringent Prism Couplers for Thin-Film Optical Waveguides," Appl. Opt. vol. 17, No. 7, pp. 1030-1037 (April 1978).
- S. P. Kuo and B. R. Cheo, "Saturation of Parametric Instabilities by Heating Effect," Bull. Am. Phys. Soc., Vol. 23, No. 7, p. 817 (September 1978).
- L. Kurz and C. S. Yoon, "Recursive Factor Analysis Methods in Feature Extraction Problems, 1977 IEEE Symp. on Infor. Theory, p. 83 (October 1977).
- L. Kurz and C. Mohwinkel, "Chain Encoding of Tabular Data in Noise Environments," 1977 IEEE Symp. on Infor. Theory, p. 89 (October 1977).
- L. Kurz, "Nonparametric Detectors Based on Partition Tests," in Nonparametric Methods in Communications: Selected Topics, edited by P. Papantoni-Kazakos and D. Kazakos, Mascel Dakker (1977).
- E. Levi, L. Birenbaum and Z. Zabar, "Concerning the Design of Inductor Synchronous Motors Fed by Current Source Inverters," IEEE Trans. on Magnetics, Vol. MAG-13, No. 5, pp. 1421-1423 (September 1977).
- E. Levi, "Magnetohydrodynamic Power Generation: Status Report," IEEE Spectrum (May 1978).
- M. C. Newstein and F. P. Mattar, "Transverse Effects Associated with the Propagation of Coherent Optical Pulses in Resonant Media," IEEE J. Quant. Electron., Vol. QE-13, pp. 507-520 (1977).
- A. A. Oliner, "IOOC '77: International Conference on 'Integrated Optics and Optical-Fiber Communication'," Scientific Bulletin, Office of Naval Research Tokyo, Vol. 2, No. 4 (October to December 1977).
- A. A. Oliner, S. T. Peng and J. P. Hsu, "New Propagation Effects for the Inverted Strip Waveguide for Millimeter Waves," International Microwave Symp. Digest, pp. 408-410, Ottawa, Canada (June 1978).
- A. A. Oliner and S. T. Peng, "Effects of Metal Overlays on 3-D Optical Waveguides," Appl. Opt., Vol. 17, No. 18, pp. 2866-2867 (September 15, 1978).

- A. Papoulis, "The Two-To-One Rule in Data Smoothing," *IEEE Trans. Inf. Theory*, Vol. IT-23, No. 5, pp. 631-633 (September 1977).
- A. Papoulis, "Generalized Sampling Expansion," *IEEE Trans. Cir. Syst.*, Vol. CAS-24, No. 11, pp. 652-654 (November 1977).
- A. Papoulis, "The Problem of Transmission Zeros in Deconvolution," *IEEE Trans. Inf. Theory*, Vol. IT-24, No. 1, pp. 126-128 (January 1978).
- A. Papoulis, "The Factorization Problem for Time-Limited Functions and Trigonometric Polynomials," *IEEE Trans. Cir. Syst.*, Vol. CAS-25, No. 1, pp. 41-45 (January 1978).
- A. Papoulis, "Identification of Systems Driven by Non-stationary Noise," *IEEE Trans. Inf. Theory*, Vol. IT-24, No. 2, pp. 240-244 (March 1978).
- S. T. Peng and A. A. Oliner, "Leakage and Resonance Effect on Strip Waveguides for Integrated Optics," *Trans. IECE of Japan*, Vol. E61, No. 3, pp. 151-154 (March 1978).
- S. T. Peng and A. A. Oliner, "Leakage and Resonance Effects on Guiding Structures for Optical and Millimeter Waves," *Proc. Seventh European Microwave Conf.*, pp. 15-19, Copenhagen, Denmark (September 1977).
- V. Shah and T. Tamir, "Brewster Phenomena in Lossy Structures," *Optics Comm.*, Vol. 23, pp. 113-117 (October 1977).
- L. Shaw, D. Sarlat and Y. Thomas, "Synthesis of Nonlinear Controllers," *Proceedings of IFAC Congress 1978*, published by Instrument Society of America, Phila., Pa.
- T. S. Sundresh, F. Cassara and H. Schachter, "Maximum A Posteriori Estimator for Suppression of Interchannel Interference in FM Receivers," *IEEE Trans. Comm.*, Vol. 25, 1480-1485 (December 1977).
- T. Tamir and S. T. Peng, "Network Methods for Integrated Optics Devices," *Proc. Int. Conf. Applic. Holography and Data Processing*, Pergamon Press, Oxford, pp. 437-446, 1977.
- T. Tamir and S. T. Peng, "Analysis and Design of Grating Couplers," *Appl. Phys.*, Vol. 14, pp. 235-254 (November 1977) (Invited Paper).
- N. Wattanapanom and L. Shaw, "Optimal Inspection Schedules for Failure Detection in a Model Where Tests Hasten Failures," *Operations Research* (to appear).
- T. Q. Yip, S. P. Kuo and B. R. Cheo, "Evolution of Parametrically Excited Instabilities in a Magneto Plasma," *Bull. Am. Phys. Soc.*, Vol. 23, No. 7, p. 799 (September 1978).

## RECENT BOOKS

- A. A. Oliner, "Acoustic Surface Waves," Vol. 24 in *Topics of Applied Physics* (Germany: Springer Verlag, 1978).
- H. Ruston, "Programming with PL/I," (New York: McGraw Hill, 1978).

- D. Baggi and M. Shooman, "Test Models: Classification and Automatic Driver Design" (POLY EE 77-040).
- S. Barone and N. Marcuvitz, "On the Theory of Plasma Turbulence II" (POLY-MRI-77-1388).
- Y. Fujimoto and E. Mishkin, "n Successive Strong Shocks Imploding a Small Sphere" (POLY EE 78-046).
- Y. Fujimoto and E. Mishkin, "Spherical Shock Implosion Self-Similar Analysis" (POLY EE 78-048).
- Y. Fujimoto and E. Mishkin, "Cylindrical and Spherical Shock Implosion. Approximate Analysis" (POLY EE 78-049).
- Y. Fujimoto and E. Mishkin, "Analysis of a Spherical and Cylindrical Shock Implosion II" (POLY EE 78-050).
- C.L. Hsu, L. Shaw and S.G. Tyan, "Reliability Applications of Multivariate Exponential Distributions," (POLY EE 77-036).
- A. Laemmel, "Study of Recursive Function Theory and Its Application to Program Complexity" (POLY EE 77-037).
- A. Laemmel, "Statistical Test Models" (POLY EE 77-041).
- E. Mishkin, "Analysis of Spherically Exploding and Imploding Shocks" (POLY EE 77-038).
- E. Mishkin and Y. Fujimoto, "The Postulate of Analyticity by Gelfand and Butler and the Analysis of a Spherically Imploding Shock" (POLY EE 78-043).
- E. Mishkin and Y. Fujimoto, "Analysis of Cylindrical Exploding and Imploding Shock Waves" (POLY EE 78-044).
- E. Mishkin and Y. Fujimoto, "Analysis of Self-Similar Shock Implosion of a Small Sphere" (POLY EE 78-045).
- M.C. Newstein, "Direct Evaluation of Diffusion Coefficient for a Prescribed Stochastic Field" (POLY MRI 77-1387).
- M.C. Newstein and N. Solimene, "Interaction of Intense Laser Rod with Metal Surface" (POLY MRI 77-1389).
- K. Park and W.T. Walter, "Theory of the Integrating Sphere for Pulse Light Sources" (POLY MRI 78-1391).
- K. Park and W.T. Walter, "Calculation of the Surface Temperature of a Metal Irradiated by a Laser Pulse" (POLY MRI 78-1392).
- G.S. Popkin, "On the Number of Tests Necessary to Verify a Computer Program" (POLY EE 77-039).
- G.S. Popkin and M.L. Shooman, "On the Number of Tests Necessary to Verify a Computer Program" (POLY EE 78-047).
- M. Shooman and H. Ruston, "Software Modelling Studies" (POLY EE 77-042).
- N. Solimene, "High Power Microwave Propagation Through the Atmosphere" (POLY MRI 78-1390).

#### DEPARTMENT OF DEFENSE

Director  
National Security Agency  
ATTN: Dr. T. J. Beahn  
Fort George G. Meade, MD 20755

Defense Documentation Center (12 copies)  
ATTN: DDC-TCA (Mrs. V. Caponio)  
Cameron Station  
Alexandria, VA 22314

Dr. George Gamota  
Acting Assistant for Research  
Deputy Under Secretary of Defense for Research  
and Engineering (Research & Advanced Technology)  
Room 3D1079, The Pentagon  
Washington, DC 20301

Mr. Leonard R. Weisberg  
Office of the Under Secretary of Defense  
for Research & Engineering/EPS  
Room 3D1079, The Pentagon  
Washington, DC 20301

Defense Advanced Research Projects Agency  
ATTN: (Dr. R. Reynolds)  
1400 Wilson Boulevard  
Arlington, VA 22209

#### DEPARTMENT OF THE ARMY

Commandant  
US Army Air Defense School  
ATTN: ATSD-T-CSM  
Fort Bliss, TX 79916

Commander  
US Army Armament R&D Command  
ATTN: DRDAR-RD  
Dover, NJ 07801

Commander  
US Army Ballistics Research Laboratory  
ATTN: DRXRD-RAD  
Aberdeen Proving Ground  
Aberdeen, MD 21005

Commandant  
US Army Command and General Staff College  
ATTN: Acquisitions, Lib. Div.  
Fort Leavenworth, KS 66027

Commander  
US Army Communication Command  
ATTN: CC-OPS-PD  
Fort Huachuca, AZ 85613

Commander  
US Army Materials and Mechanics Research Center  
ATTN: Chief, Materials Sciences Division  
Watertown, MA 02172

Commander  
US Army Materiel Development  
and Readiness Command  
ATTN: Technical Library, Rm. 7S 35  
5001 Eisenhower Avenue  
Alexandria, VA 22333

Commander  
US Army Missile R&D Command  
ATTN: Chief, Document Section  
Redstone Arsenal, AL 35809

Commander  
US Army Satellite Communications Agency  
Fort Monmouth, NJ 07703

Commander  
US Army Security Agency  
ATTN: IARD-T  
Arlington Hall Station  
Arlington, VA 22212

Project Manager  
Army Tactical Data Systems  
EAI Building  
West Long Branch, NJ 07764

Commander  
Atmospheric Sciences Laboratory (ERADCOM)  
ATTN: DELAS-BL-RD  
White Sands Missile Range, NM 88002

Director  
US Army Electronics R&D Command  
Night Vision & Electro-Optics Labs  
ATTN: Dr. Ray Balcerak  
Fort Belvoir, VA 22060

Commander  
US Army Communications R&D Command  
ATTN: DRDCO-COM-C (Dr. Herbert S. Bennett)  
Fort Monmouth, NJ 07703

Commander  
US Army Research Office  
ATTN: DRXRO-MA (Dr. Paul Boggs)  
P. O. Box 12211  
Research Triangle Park, NC 27709

Commander  
US Army Missile R&D Command  
Physical Sciences Directorate  
ATTN: DRDMI-TRD (Dr. Charles Bowden)  
Redstone Arsenal, AL 35809

Director  
TRI-TAC  
ATTN: TT-AD (Mrs. Briller)  
Fort Monmouth, NJ 07703

Commander  
US Army Missile R&D Command  
Advanced Sensors Directorate  
ATTN: DRDMI-TER (Dr. Don Burlage)  
Redstone Arsenal, AL 35809

Commander  
US Army Electronics R&D Command  
Night Vision & Electro-Optics Labs  
ATTN: DELNV (Dr. Rudolf G. Buser)  
Fort Monmouth, NJ 07703

Director  
US Army Electronics R&D Command  
Night Vision & Electro-Optics Labs  
ATTN: Mr. John Dehns  
Fort Belvoir, VA 22060

Director  
US Army Electronics R&D Command  
Night Vision & Electro-Optics Labs  
ATTN: Dr. William Ealy  
Fort Belvoir, VA 22060

Director  
US Army Electronics R&D Command  
ATTN: DELEV (Electronic Warfare Laboratory)  
White Sands Missile Range, NM 88002

Executive Secretary, TAC/JSEP  
US Army Research Office  
P. O. Box 12211  
Research Triangle Park, NC 27709

Commander  
US Army Missile R&D Command  
Physical Sciences Directorate  
ATTN: DRDMI-TER (Dr. Michael D. Fahey)  
Redstone Arsenal, AL 35809

Commander  
US Army Missile R&D Command  
Physical Sciences Directorate  
ATTN: DRDMI-TRO (Dr. William L. Gamble)  
Redstone Arsenal, AL 35809

Commander  
White Sands Missile Range  
ATTN: STEWS-ID-SR (Dr. Al L. Gilbert)  
White Sands Missile Range, NM 88002

Project Manager  
Ballistic Missile Defense Program Office  
ATTN: DACS-DMP (Mr. A. Gold)  
1300 Wilson Blvd.  
Arlington, VA 22209

Commander  
US Army Communications R&D Command  
ATTN: CENTACS (Dr. David Haratz)  
Fort Monmouth, NJ 07703

Commander  
Harry Diamond Laboratories  
ATTN: Mr. John E. Rosenberg  
2800 Powder Mill Road  
Adelphi, MD 20783

RQDA (DAMA-ARZ-A)  
Washington, DC 20310

Commander  
US Army Electronics R&D Command  
ATTN: DELET-E (Dr. Jack A. Kohn)  
Fort Monmouth, NJ 07703

Commander  
US Army Electronics Technology & Devices Lab  
ATTN: DELET-EN (Dr. S. Kroenenberg)  
Fort Monmouth, NJ 07703

Commander  
US Army Communications R&D Command  
ATTN: CENTACS (Mr. R. Kulinyi)  
Fort Monmouth, NJ 07703

Commander  
US Army Communications R&D Command  
ATTN: DRDCO-TCS-BG (Dr. E. Lieblein)  
Fort Monmouth, NJ 07703

Commander  
US Army Electronics Technology and Devices Lab  
ATTN: DELET-HM (Mr. N. Lipetz)  
Fort Monmouth, NJ 07703

Director  
US Army Electronics R&D Command  
Night Vision & Electro-Optics Labs  
ATTN: Dr. Randy Longshore  
Fort Belvoir, VA 22060

Commander  
US Army Electronics R&D Command  
ATTN: DRDEL-CT (Dr. W. S. McAfee)  
2800 Powder Mill Road  
Adelphi, MD 20783

Commander  
US Army Research Office  
ATTN: DRXRO-EL (Dr. James Mink)  
P. O. Box 12211  
Research Triangle Park, NC 27709

Director  
US Army Electronics R&D Command  
Night Vision Laboratory  
ATTN: DELNV  
Fort Belvoir, VA 22060

COL Robert Noco  
Senior Standardization Representative  
US Army Standardization Group, Canada  
Canadian Force Headquarters  
Ottawa, Ontario, Canada K1A 0K2

Commander  
Harry Diamond Laboratories  
ATTN: Dr. Robert Oswald, Jr.  
2800 Powder Mill Road  
Adelphi, MD 20783

Commander  
US Army Communications R&D Command  
ATTN: CENTACS (Dr. D. C. Pearce)  
Fort Monmouth, NJ 07703

Director  
US Army Electronics R&D Command  
Night Vision & Electro-Optics Labs  
ATTN: DELNV-ED (Dr. John Pollard)  
Fort Belvoir, VA 22060

Commander  
US Army Research Office  
ATTN: DRXRO-EL (Dr. William A. Sander)  
P. O. Box 12211  
Research Triangle Park, NC 27709

Commander  
US Army Communications R&D Command  
ATTN: DRDCO-COM-RH-1 (Dr. Felix Schwering)  
Fort Monmouth, NJ 07703

Commander  
US Army Electronics Technology and Devices Lab  
ATTN: DELET-1 (Dr. C. G. Thornton)  
Fort Monmouth, NJ 07703

U. S. Army Research Office (3 copies)  
ATTN: Library  
P. O. Box 12211  
Research Triangle Park, NC 27709

Director  
Division of Neuropsychiatry  
Walter Reed Army Institute of Research  
Washington, DC 20012

Commander  
US ARRAADCOM  
ATTN: DRDAR-SCF-CC (Dr. N. Coleman)  
Dover, NJ 07801

#### DEPARTMENT OF THE AIR FORCE

Mr. Robert Barrett  
RADC/ES  
Hanscom AFB, MA 01731

Dr. Carl E. Baum  
AFVL (ES)  
Kirtland AFB, NM 87117

Dr. E. Champagne  
AFAL/DH  
Wright-Patterson AFB, OH 45433

Dr. R. F. Dolan  
RADC/ESR  
Hanscom AFB, MA 01731

Mr. W. Edwards  
AFAL/DH  
Wright-Patterson AFB, OH 45433

Professor R. E. Fontana  
Head Dept. of Electrical Eng.  
AFIT/ENE  
Wright-Patterson AFB, OH 45433

Dr. Alan Garscadden  
AFAPL/POD  
Wright-Patterson AFB, OH 45433

USAF European Office of Aerospace Research  
ATTN: Major J. Gorrell  
Box 14, FPO, New York 09510

LTC Richard J. Cowen  
Department of Electrical Engineering  
USAF Academy, CO 80840

Mr. Murray Kesselman (ISCA)  
Rome Air Development Center  
Griffiss AFB, NY 13441

Dr. G. Knausenberger  
Air Force Member, TAC  
Air Force Office of Scientific Research  
(AFSC) AFOSR/NE  
Bolling Air Force Base, DC 20332

COL R. V. Gomez  
Air Force Member, TAC  
Air Force Office of Scientific Research  
(AFSC) AFOSR/NE  
Bolling Air Force Base, DC 20332

Mr. R. D. Larson  
AFAL/DHR  
Wright-Patterson AFB, OH 45433

Dr. Edward Althuler  
RADC/EEP  
Hanscom AFB, MA 01731

Mr. John Mottamith (NCI)  
HQ ESD (AFSC)  
Hanscom AFB, MA 01731

Dr. Richard Picard  
RADC/ETSL  
Hanscom AFB, MA 01731

Dr. J. Ryles  
Chief Scientist  
AFAL/CA  
Wright-Patterson AFB, OH 45433

Dr. Allan Schell  
RADC/EE  
Hanscom AFB, MA 01731

Mr. H. E. Webb, Jr. (ISCP)  
Rome Air Development Center  
Griffiss AFB, NY 13441

Dr. R. Kelley  
Air Force Office of Scientific Research  
(AFSC) AFOSR/NP  
Bolling Air Force Base, DC 20332

LTC G. McKemie  
Air Force Office of Scientific Research  
(AFSC) AFOSR/NW  
Bolling Air Force Base, DC 20332

#### DEPARTMENT OF THE NAVY

Office of Naval Research  
800 North Quincy Street  
Arlington, VA 22217  
Attn: Codes 220/221  
427  
432

Naval Research Laboratory  
4555 Overlook Avenue, SW  
Washington, DC 20375  
Attn: Codes 1405 - Dr. S. Teitler  
2627 - Mrs. D. Folen  
5200 - A. Brodzinsky  
5210 - J. E. Davey  
5270 - B. D. McCombe  
5403 - J. E. Shore  
5464/5410 - J. R. Davis  
5510 - W. L. Faust  
7701 - J. D. Brown

Director  
Office of Naval Research Branch Office  
495 Summer Street  
Boston, MA 02210

Director  
Office of Naval Research  
New York Area Office  
715 Broadway, 5th Floor  
New York, NY 10003

Director  
Office of Naval Research Branch Office  
536 South Clark Street  
Chicago, IL 60605

Director  
Office of Naval Research Branch Office  
1030 East Green Street  
Pasadena, CA 91101

Office of Naval Research  
San Francisco Area Office  
760 Market Street, Room 447  
San Francisco, CA 94102

Naval Surface Weapons Center  
Attn: Technical Library  
Code DX-21  
Dahlgren, VA 22448

Dr. J. H. Mills, Jr.  
Naval Surface Weapons Center  
Code DF  
Dahlgren, VA 22448

Naval Air Development Center  
Johnsville  
Warminster, PA 18974  
Attn: Codes 01 - Dr. R. Lobb  
202 - T. Shoppie  
Technical Library

Dr. Gernot M. R. Winkler  
Director, Time Service  
U. S. Naval Observatory  
Mass. Avenue at 34th Street, NW  
Washington, DC 20390

Dr. G. Gould  
Technical Director  
Naval Coastal Systems Laboratory  
Panama City, FL 32401

Dr. W. A. VonWinkle  
Associate Technical Director for Technology  
Naval Underwater Systems Center  
New London, CT 06320

Naval Underwater Systems Center  
Attn: J. Merrill  
Newport, RI 02840

Technical Director  
Naval Underwater Systems Center  
New London, CT 06320

Naval Research Laboratory  
Underwater Sound Reference Division  
Technical Library  
P. O. Box 8337  
Orlando, FL 32806

Naval Ocean Systems Center  
San Diego, CA 92152  
Attn: Codes 01 - H. L. Blood  
015 - P. C. Fletcher  
9102 - W. J. Dejka  
922 - H. M. Wiedner  
532 - J. H. Richter

Naval Weapons Center  
China Lake, CA 93555  
Attn: Codes 601 - P. C. Essig  
5515 - M. H. Ritchie

Donald E. Kirk  
Professor & Chairman, Electronic Engineering  
Sp-304  
Naval Postgraduate School  
Monterey, CA 93940

Mr. J. C. French  
National Bureau of Standards  
Electronics Technology Division  
Washington, DC 20234

Harris B. Stone  
Office of Research, Development, Test & Evaluation  
NRP-987  
The Pentagon, Room 50760  
Washington, DC 20350

Dr. A. L. Slatkokey  
Code RD-1  
Headquarters Marine Corps  
Washington, DC 20380

Dr. H. J. Mueller  
Naval Air Systems Command  
Code 310  
JP #1  
1411 Jefferson Davis Hwy.  
Arlington, VA 20360

Mr. Larry Sumney  
Naval Electronics Systems Command  
Code 03R  
NC #1  
2511 Jefferson Davis Hwy.  
Arlington, VA 20360

Naval Sea Systems Command  
NC #3  
2531 Jefferson Davis Hwy.  
Arlington, VA 20362  
Attn: Code 03C - J. H. Nuth

Officer in Charge  
Carderock Laboratory  
Code 322.1 - Technical Library  
Code 1B - G. H. Gleissner  
David Taylor Naval Ship Research & Development Center  
Bethesda, MD 20084

Naval Surface Weapons Center  
White Oak  
Silver Spring, MD 20910  
Attn: Codes VX-40 - Technical Library  
WR-303 - R. S. Allgaier  
WR-34 - H. R. Riedl

#### OTHER GOVERNMENT AGENCIES

Dr. Howard W. Etzel  
Deputy Director  
Division of Materials Research  
National Science Foundation  
1800 G Street  
Washington, DC 20550

Mr. J. C. French  
National Bureau of Standards  
Electronics Technology Division  
Washington, DC 20234

Dr. Jay Harris  
Program Director  
Devices and Waves Program  
National Science Foundation  
1800 G Street  
Washington, DC 20550

Los Alamos Scientific Laboratory  
ATTN: Reports Library  
P. O. Box 1663  
Los Alamos, NM 87544

Dr. Dean Mitchell  
Program Director, Solid-State Physics  
Division of Materials Research  
National Science Foundation  
1800 G Street  
Washington, DC 20550

Mr. F. C. Schwenk, RD-T  
National Aeronautics & Space Administration  
Washington, DC 20546

M. Zane Thornton  
Deputy Director Institute for  
Computer Sciences and Technology  
National Bureau of Standards  
Washington, DC 20234

Head, Electrical Sciences & Analysis Section  
National Science Foundation  
1800 G Street, NW  
Washington, DC 20550

#### NON-GOVERNMENT AGENCIES

Director  
Columbia Radiation Laboratory  
Columbia University  
538 West 120th Street  
New York, NY 10027

Director  
Coordinated Science Laboratory  
University of Illinois  
Urbana, IL 61801

Director  
Division of Engineering and  
Applied Physics  
Harvard University  
Pierce Hall  
Cambridge, MA 02138

Director  
Electronics Research Center  
The University of Texas  
P. O. Box 7728  
Austin, TX 78712

Director  
Electronics Research Laboratory  
University of California  
Berkeley, CA 94720

Director  
Electronics Sciences Laboratory  
University of Southern California  
Los Angeles, CA 90007

Director  
Microwave Research Institute  
Polytechnic Institute of New York  
333 Jay Street  
Brooklyn, NY 11201

Director  
Research Laboratory of Electronics  
Massachusetts Institute of Technology  
Cambridge, MA 02139

Director  
Stanford Electronics Laboratory  
Stanford University  
Stanford, CA 94305

Director  
Stanford Ginton Laboratory  
Stanford University  
Stanford, CA 94305

Dr. Lester Eastman  
School of Electrical Engineering  
Cornell University  
Ithaca, NY 14850

Chairman  
Department of Electrical Engineering  
Georgia Institute of Technology  
Atlanta, GA 30332

Dr. Carlton Walter  
ElectroScience Laboratory  
The Ohio State University  
Columbus, OH 43212

Dr. Richard Seeks  
Department of Electrical Engineering  
Texas Tech University  
Lubbock, TX 79409

Dr. Roy Gould  
Executive Officer for Applied Physics  
California Institute of Technology  
Pasadena, CA 91125

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM	
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle)	5. TYPE OF REPORT & PERIOD COVERED	6. PERFORMING ORG. REPORT NUMBER	
PROGRESS REPORT No. 43 TO THE JOINT SERVICES TECHNICAL ADVISORY COMMITTEE.	Scientific / Interim / Rept. 13	POLY-MRI-452.43-78	
7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)	9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Arthur A. Oliner	F44620-78-C-0074, F 44620-74-C-0056	Polytechnic Institute of New York Microwave Research Institute 333 Jay Street, Brooklyn, NY 11201	Project No. 4751 61102F 681305
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE	13. NUMBER OF PAGES	14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Joint Services Electronic Program through Dir. of Electronic and Solid State Sciences Air Force Office of Scientific Research (NE) Bolling AFB, Washington, DC 20332	November 1978	566	12) 568p.
15. SECURITY CLASS. (of this report)		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
UNCLASSIFIED			
16. DISTRIBUTION STATEMENT (of this Report)			
1. This document has been approved for public release and sale; its distribution is unlimited.			
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)			
18. SUPPLEMENTARY NOTES			
TECH, OTHER			
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)			
Acoustic Correlators		Control Theory	
Antennas		Data Processing	
Communications		Electric Power Engineering	
Computer Communications		Electromagnetics	
Computers		Electronics	
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)			
<p>→ This report is the forty-third in a series of progress reports to the Joint Services Technical Advisory Committee since inception of the Joint Services Electronics Program at the Polytechnic Institute of Brooklyn in July 1955. The report, now being issued annually, summarizes research accomplished under the aegis of the Microwave Research Institute and reflects the impact of the Joint Services Electronics Program on the research activities of faculty and students of the Institute. The program covers a broad spectrum ranging from basic theoretical physics, mathematics, and engineering, to experimen-</p>			

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

557

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

409138

1/B

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

19.

Electrophysics  
Energy Conversion  
Micro-Electronics  
Microwave Acoustics  
Network Theory  
Optics

Quantum Electronics  
Software Reliability and Engineering  
Solid State and Materials  
Systems Control  
Waveguide Techniques  
Wave-Matter Interactions

20.

Final investigations involving basic measurements, development of devices, and materials.

Each activity reports in summary fashion on specific results obtained during the report period, 15 September 1977 through 14 September 1978, with individual acknowledgement of the sponsorship which has contributed to the reported work. The report is organized into two major divisions. The first, Electrophysics, includes the topics of: Electromagnetics; Acoustics; Optics; Quantum Electronics; Solid State and Materials; Wave-Matter Interactions; and Electric Power Engineering. The second, Systems, includes the topics of: Communications; Computer and Computer-Communications Networks; Safety, Reliability and Software Engineering; Systems, Control and Networks; and Data Processing.

A

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)